NOTAS DE TÉCNICAS DE MUESTREO

Luis Valdivieso Serrano

DEPARTAMENTO ACADÉMICO DE **CIENCIAS**

PUCP

NOTAS DE TÉCNICAS DE MUESTREO

Luis Valdivieso Serrano

DEPARTAMENTO
ACADÉMICO DE
CIENCIAS

PUCP

Autor
Luis Valdivieso Serrano

@Pontificia Universidad Católica del Perú
Departamento Académico de Ciencias
Sección Matemáticas
Av. Universitaria 1801, San Miguel
Teléfono: 6262000
Correo electrónico: publicacionesdac@pucp.edu.pe

Notas de Técnicas de Muestreo

Lima, Departamento Académico de Ciencias -
Sección Matemática, 2020

Diseño y diagramación: Elit León
Calle Santa Francisca Romana 395, Lima
Teléfono: 6571260
elit.leon@gmail.com

Primera edición digital: diciembre de 2020

Publicación disponible en: https://departamento.pucp.edu.pe/ciencias/investigaciones-y-publicaciones/publicaciones-del-departamento/

ISBN: 978-612-47757-1-0
Hecho el Depósito Legal en la Biblioteca Nacional del Perú: 2020-09987

Presentación

Este texto, que fue inicialmente redactado como material de apoyo para los estudiantes de la maestría en Estadística de la Pontificia Universidad Católica del Perú, ofrece una introducción al estudio de las principales técnicas de muestreo probabilístico.

Si bien en la literatura existen varios textos clásicos sobre muestreo como el de Cochran (1977), Mendenhall et al. (2007) y Lohr (2000) y más avanzados como el de Tillé (2006) y Lumley (2010), falta todavía, a mi humilde opinión, un texto de nivel intermedio que integre estos enfoques y que a su vez incluya más aplicaciones a datos reales de dominio público. Estas notas pretenden cubrir tal vacío presentando no solo las técnicas de muestreo probabilístico clásicas, sino también tópicos de muestreo complejo y una implementación computacional que actúe transversalmente a lo largo de los diferentes temas del curso. Para ello usaremos principalmente los paquetes survey y sampling escritos en el software libre R. Información sobre estos se puede consultar, respectivamente, en los siguientes enlaces:

`http://cran.r-project.org/web/packages/survey/survey.pdf`

`https://cran.r-project.org/web/packages/sampling/sampling.pdf`

o en los textos de Lumley (2010) y Tillé (2006). Otra excelente referencia en el espíritu de estas notas, y que incluye al paquete PracTools de R, es Valliant et al. (2013).

El texto está dividido en cinco capítulos. En el primer capítulo introducimos algunos conceptos básicos de estadística y ponemos énfasis en la diferencia que existe entre los enfoques basados en el modelo y en el diseño. En el segundo capítulo presentamos la teoría del muestreo aleatorio simple (MAS) e introducimos aquí no solo los conceptos teóricos pertinentes, sino también su implementación computacional y aplicación a datos reales. En el tercer capítulo definimos el muestreo aleatorio estratificado como el agregado de un MAS aplicado a subconjuntos relativamente homogéneos de la población, a los cuales denominaremos estratos. En el capítulo cuatro abordamos el muestreo por conglomerados, el cual es quizás el esquema clásico más utilizado para grandes poblaciones. A diferencia del diseño anterior, este esquema resulta ser más eficiente cuando los subconjuntos de la población (que denominaremos conglomerados) muestran una marcada heterogeneidad en su interior pero gran similitud entre ellos. Un tema central y unificador en este capítulo será el estudio de los estimadores de Horvitz-Thompson para totales en diseños de conglomerados de una o

más etapas con probabilidades de selección no siempre constantes. De este se derivan casi todos los esquemas anteriores, como el de conglomerados de una etapa y su caso particular el muestreo sistemático. En el último capítulo nos dedicamos al estudio de muestras complejas. Estas se originan cuando debido a la configuración y al tamaño de la población en estudio se hace necesario restringir o combinar dos o más técnicas, ya sea que cada selección se haga con igual probabilidad o no. Aquí nos interesará no solo obtener estimaciones puntuales de los parámetros de interés, al expandir apropiadamente la muestra a la población, sino fundamentalmente estimar la variabilidad de las estimaciones. Para ello discutiremos diversas técnicas como la linealización y el remuestreo y nos apoyaremos, al igual que en los capítulos anteriores, en los paquetes survey y sampling de R. Este capítulo brindará también una introducción al análisis estadístico bajo muestras complejas. Como ilustración, veremos aquí el análisis de datos categóricos, el de regresión y los contrastes de hipótesis para una, dos o más poblaciones. El capítulo incluye algunos diseños muestrales y sus correspondiente análisis para las bases de datos introducidas en el curso.

El texto se complementa con diversos ejercicios propuestos y algunas sugerencias o soluciones a estos en un anexo final. Tales ejercicios son de nivel teórico y práctico y se usan, en muchos de ellos, bases de datos de dominio público tanto locales como foráneas.

Dr. Luis Valdivieso

Índice general

Capítulo 1

Introducción

1.1. Enfoques basados en el diseño y el modelo

Supongamos que un banco busca estimar el ahorro medio que las familias de un distrito planifican para un mes. Sea y la variable (estadística) que asigna a cada familia del distrito este monto de ahorro en soles. Naturalmente, si aquí se hace un censo en el que se pregunte y averigue (con fortuna) sobre los ahorros de las N familias del distrito, uno obtendrá N números $y_1, y_2, \ldots, y_N$ y el ahorro medio de interés será:

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

Desafortunadamente, el banco no puede hacer un censo, y por ello planifica realizar un muestreo probabilístico seleccionando al azar, y por simplicidad con reemplazamiento, una por una a las familias del padrón de la municipalidad hasta un número $n < N$. Note que bajo este esquema toda familia tiene la misma probabilidad de ser escogida. Al término del estudio, el banco obtendrá la muestra

$$Y_1, Y_2, \ldots, Y_n, \tag{1.1}$$

donde Y_i denota el valor (aleatorio) que podría tomar la variable estadística y en la i-ésima selección de la muestra. Realizadas las observaciones, el ahorro medio mensual de las familias del distrito podrá estimarse mediante la media aritmética de estos valores. Note aquí que la aleatoriedad es introducida por el esquema de selección en el diseño de la muestra. Así, podríamos escribir indistintamente la variable aleatoria correspondiente a la estimación anterior como

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{o} \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{N} y_i\delta_i, \tag{1.2}$$

siendo δ_i una variable aleatoria con distribución binomial de parámetros n y probabilidad $\frac{1}{N}$ que denota el número de veces que la i-ésima familia del distrito es seleccionada en la muestra.

Estadísticamente, (1.2) es un buen estimador de μ_N. Como podemos ver, su valor esperado o media es precisamente el parámetro que buscamos; es decir, $\bar{Y}$ es un estimador insesgado de μ_N:

$$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{N} y_i E(\delta_i) = \frac{1}{n}\sum_{i=1}^{N} y_i \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu_N$$

El enfoque hasta aquí comentado se denomina enfoque basado en el diseño. Un lector perspicaz podría preguntarse por qué este difiere del esquema clásico de inferencia en el que uno simplemente asume una distribución o "superpoblación" para el ahorro Y de las familias del distrito, digamos normal con media μ y varianza σ^2 y, por tanto, estima μ (que es la cantidad que el banco quiere) al tomarse una muestra aleatoria $Y_1, Y_2, \ldots, Y_n$ de Y y considerarse el estimador

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

La respuesta a esta interrogante no es tan directa. El enfoque clásico comentado, que se basa en el modelo normal, difiere del que se basa en el diseño en el sentido que los parámetros poblacionales μ y μ_N son por naturaleza distintos, a menos que la población sea infinita y el modelo esté bien especificado. En efecto, uno puede integrar ambos enfoques pensando que si la población fuese hipotéticamente grande ($N \to \infty$), entonces la distribución empírica de los números $y_1, y_2, \ldots, y_N$ (piense por simplicidad en el polígono de frecuencias relativas del histograma de estos datos) debería converger (si el modelo es correcto) hacia la curva normal. Luego podríamos pensar en la colección dada por (1.1) como una muestra aleatoria de la variable aleatoria Y. En la realidad, las poblaciones son finitas; por ello si el interés radica en estudiar la variable y, uno podría asumir que esta población es a su vez una muestra representativa de la superpoblación.

Observe que en un modelo basado en el diseño, a diferencia que en el de su par basado en el modelo, la distribución de Y es irrelevante, a menos que, como precisamos, uno tenga interés y tenga sentido analizar cuestiones asintóticas. Desde un punto de vista práctico, el enfoque basado en el diseño nos será más útil para estudiar poblaciones finitas; mientras que el enfoque basado en el modelo lo será para el estudio de poblaciones infinitas o muy grandes.

Resumiendo, en un enfoque basado en el diseño, la aleatoriedad es introducida por el esquema de selección de las unidades en la muestra, y la población finita de tamaño N sobre la que se mide una o más características, como y, es fija e invariable. Por otro lado, en un enfoque basado en el modelo, la aleatoriedad es introducida por el modelo de distribución que se asigne a la variable de interés. Así, ella define una superpoblación con un número muy grande o infinito de posibles valores para esta variable, y los valores que ella toma en la población finita de tamaño N conforman tan solo un subconjunto que se espera sea representativo de esta superpoblación.

1.2. Estimadores puntuales y por intervalos

Al margen del enfoque o diseño muestral utilizado, existen tres características primordiales que uno debe tomar en cuenta en todo estudio inferencial. Estas son: el tamaño de la muestra que se utilizará, el nivel de confianza y el error de estimación. Todos estos conceptos están íntimamente ligados a la teoría de la estimación puntual y por intervalos, puntos que revisamos brevemente antes de presentar los principales tipos de muestreo probabilístico.

Sea X una variable aleatoria (v.a.) cuya distribución depende de un parámetro poblacional desconocido θ. Dada una muestra aleatoria (m.a.) de tamaño n de X; vale decir, una colección $X_1, X_2, \ldots, X_n$ de n v.a. independientes y con la misma distribución que X, es de interés obtener un estimador $\hat{\theta}_n = g(X_1, X_2, \ldots, X_n)$ de θ. Por definición, este estimador puede ser cualquier estadística (función de la m.a.), pero es claro que nos interesarán estimadores buenos en el sentido, que de observarse la muestra, podamos garantizar que el valor observado $g(x_1, x_2, \ldots, x_n)$ de $\hat{\theta}_n$, al que llamaremos una estimación, se ubique cerca a θ. Dado que no conocemos θ, esta cercanía debe evaluarse por métodos probabilísticos. En general, un buen estimador, $\hat{\theta}_n$ de θ, debe verificar en lo posible las siguientes tres propiedades básicas:

- $\hat{\theta}_n$ debe ser un estimador insesgado; i.e, $E(\hat{\theta}_n) = \theta$
- $\hat{\theta}_n$ debe ser eficiente; i.e, debe tener varianza pequeña, por lo usual mínima bajo una clase de estimadores insesgados
- $\hat{\theta}_n$ debe ser consistente; i.e, $\hat{\theta}_n \xrightarrow{P} \theta$, conforme $n \to \infty$

Si bien el error estándar de estimación de $\hat{\theta}_n$, definido como la desviación estándar de $\hat{\theta}_n$, podría resumir la calidad del estimador, la estimación puntual no nos brinda información de cuán cerca o lejos se pueda encontrar la estimación de $\hat{\theta}_n$ de θ. Por tal motivo, surge la llamada estimación por intervalos.

Un intervalo de confianza (IC) al $100(1-\alpha)\,\%$ para un parámetro poblacional θ de una v.a. X es un intervalo con estadísticas L_1 y L_2 en sus extremos ($IC = [L_1, L_2]$), tal que

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha.$$

Una técnica para obtener un IC es utilizar alguna variable pivote de distribución conocida que dependa de la m.a. y de solo θ como valor desconocido. Por ejemplo, si deseamos estimar la media de una v.a. $X \sim N(\mu, \sigma^2)$ con varianza conocida, podríamos utilizar como variable pivote a

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Tomando luego dos valores de esta v.a., cuyas áreas en las colas sean iguales a $\frac{\alpha}{2}$ (¿por qué?), obtendremos el siguiente intervalo de confianza al $100(1-\alpha)\,\%$ para μ:

$$IC = [\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}] \,.$$

Cabe destacar que, gracias al teorema del límite central (TLC), este IC es aún válido para la media de cualquier distribución, siempre que n sea lo suficientemente grande y se tenga, de no conocerse σ, una estimación consistente de esta desviación estándar.

Otro parámetro recurrente en diversas aplicaciones lo constituye la proporción p de elementos en la población que comparten cierta característica. A fin de obtener un intervalo de confianza aproximado al $100(1-\alpha)\,\%$ para p, tomemos al azar n elementos de la población física y consideremos las v.a. X_i definidas como 1 si es que en la i-ésima selección se encuentra un elemento con la característica buscada y como 0 en caso contrario. Note que los elementos de esta muestra solo podrán garantizarse distintos si es que la muestra se toma sin reemplazamiento. Esto ocasiona que las variables $X_1, X_2, \ldots, X_n$ no sean más independientes; sin embargo, si el tamaño N de la población es grande o infinito, se podría garantizar una casi independencia (veremos un tratamiento más formal en el capítulo 2). En la práctica, si N es grande, estas variables se consideran independientes, por lo que la distribución de $X = \sum_{i=1}^{n} X_i$, que representa al número de elementos en la muestra que comparten la característica buscada, puede asumirse que tiene aproximadamente una distribución binomial de parámetros n y p. Más aún, si n es grande, podremos utilizar la aproximación de la distribución binomial por la normal y usar:

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1),$$

con $\bar{p} = \frac{X}{n}$, como variable pivote para la construcción del IC para p. En efecto, tomando simétricamente valores $-z_{1-\frac{\alpha}{2}}$ y $z_{1-\frac{\alpha}{2}}$ en la tabla normal estándar, podemos afirmar que

$$P(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha.$$

Con el fin de despejar p en esta expresión, podemos considerar la probabilidad equivalente

$$P(|\frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}|^2 \leq z^2_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

o

$$P(p^2(1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}) - p(2\bar{p} + \frac{z^2_{1-\frac{\alpha}{2}}}{n}) + \bar{p}^2 \leq 0) = 1 - \alpha.$$

Esta probabilidad puede escribirse como

$$P((p - p_1)(p - p_2) \leq 0) = 1 - \alpha,$$

donde p_1 y p_2 constituyen las raíces de la ecuación cuadrática asociada a la inecuación anterior, las cuales vienen explícitamente dadas por

$$p_1 = \frac{2\bar{p} + \frac{z^2_{1-\frac{\alpha}{2}}}{n} - \sqrt{(2\bar{p} + \frac{z^2_{1-\frac{\alpha}{2}}}{n})^2 - 4\bar{p}^2(1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n})}}{2(1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n})} = \frac{\bar{p} + \frac{z^2_{1-\frac{\alpha}{2}}}{2n} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\bar{p}(1-\bar{p}) + \frac{z^4_{1-\frac{\alpha}{2}}}{n^2}}}{1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}}$$

y

$$p_2 = \frac{2\bar{p} + \frac{z^2_{1-\frac{\alpha}{2}}}{n} + \sqrt{(2\bar{p} + \frac{z^2_{1-\frac{\alpha}{2}}}{n})^2 - 4\bar{p}^2(1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n})}}{2(1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n})} = \frac{\bar{p} + \frac{z^2_{1-\frac{\alpha}{2}}}{2n} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\bar{p}(1-\bar{p}) + \frac{z^4_{1-\frac{\alpha}{2}}}{n^2}}}{1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}}.$$

Del análisis de los signos de la inecuación al interior de la probabilidad anterior se sigue que

$$P(p_1 \leq p \leq p_2) = 1 - \alpha$$

y, por tanto, $[p_1\ ,\ p_2]$ es un intervalo de confianza al $100(1-\alpha)\,\%$ para p . Este se conoce como el intervalo de Wilson. Si, por otro lado, para simplificar despreciamos aquí al término $\frac{z^2_{1-\frac{\alpha}{2}}}{n}$, por ser este pequeño cuando n es grande, obtendremos para p el $IC = [p_1, p_2]$ al $100(1-\alpha)\,\%$ siguiente:

$$IC = [\bar{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\ ,\ \bar{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}].$$

Este se conoce como el intervalo de Wald para p.

1.3. Distribuciones importantes asociadas al muestreo

Aparte de la muy conocida distribución normal, en el curso requeriremos las formas tanto univariadas como multivariadas de las distribuciones binomial e hipergeométrica. Estas las asociaremos luego al contexto de los muestreos con y sin reemplazamiento, respectivamente.

1.3.1. La distribución binomial

Consideremos un experimento aleatorio sencillo, llamado de Bernoulli, que tiene solo dos posibles resultados: E (de éxito) y F (de fracaso). Sea $p = P(E)$ la probabilidad de que ocurra un éxito. Si repetimos este experimento n veces de manera independiente y definimos la variable aleatoria

$$X = \quad \text{Número de éxitos en los } n \text{ experimentos independientes de Bernoulli,}$$

entonces diremos que X es una v.a. con distribución binomial de parámetros n y p, y la denotaremos por $X \sim B(n,p)$.

Proposición 1.1. *Si $X \sim B(n,p)$, entonces:*

a) *La función de probabilidad de X viene dada por*

$$P_X(x) = P(X = x) = \begin{cases} C^n_x p^x (1-p)^{n-x} & \text{si } x = 0, 1, 2, ..., n \\ 0 & \text{en otro caso} \end{cases}$$

b) $E(X) = np$

c) $V(X) = np(1-p)$

Demostración: **a)** *Note que el conjunto de posibles valores que X pueda tomar (rango de X) es $R_X = \{0, 1, 2, \cdots, n\}$, ya que puede ocurrir que nunca se presente el éxito, en cuyo caso X valdrá 0; ocurra una sola vez, en cuyo caso X valdrá 1, y así sucesivamente hasta el caso extremo en que el éxito siempre esté presente, en cuyo caso X será n. Ahora bien, que el éxito se presente en x oportunidades específicas y que el fracaso ocurra en los $(n-x)$ experimentos restantes tiene la siguiente probabilidad:*

$$\overbrace{(p.p.\cdots.p)}^{x \quad términos} \quad \overbrace{(1-p)(1-p)\cdots(1-p)}^{(n-x) \quad términos} = p^x(1-p)^{n-x}.$$

Dado que en total hay C_x^n casos como este (piense en el número total de x posiciones que se podrían escoger de las n para que en ellas ocurra el éxito), se tiene que $P(X = x) = C_x^n p^x (1-p)^{n-x}$, siendo x un valor cualesquiera de $R_X = \{0, 1, 2, \cdots, n\}$.
b) *Haciendo en la sumatoria de abajo el cambio de variable $k = x - 1$, se tiene que*

$$E(X) = \sum_{x=0}^{n} x C_x^n p^x (1-p)^{n-x} = n \sum_{x=1}^{n} C_{x-1}^{n-1} p^x (1-p)^{n-x}$$

$$= np \sum_{k=0}^{n-1} C_k^{n-1} p^k (1-p)^{n-1-k} = np(p + 1 - p)^{n-1} = np.$$

c) *De manera similar, se cumple que*

$$E(X^2) = \sum_{x=0}^{n} x^2 C_x^n p^x (1-p)^{n-x} = np \sum_{k=0}^{n-1} (k+1) C_k^{n-1} p^k (1-p)^{n-1-k}$$

$$= np((n-1)p + 1) = n(n-1)p^2 + np.$$

Por tanto, $V(X) = E(X^2) - E(X)^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p)$. ■

1.3.2. La distribución multinomial

Esta es la extensión multivariada de la distribución anterior. Para describirla, consideremos un experimento aleatorio cuyos resultados pueden caer en cualquiera de k categorías excluyentes y exhaustivas $C_1, C_2, \ldots, C_k$, con probabilidades respectivas $p_1, p_2, \ldots, p_k$ que satisfacen $\sum_{i=1}^{k} p_i = 1$. Si este experimento se repite de manera independiente n veces y se definen las variables aleatorias

$$X_i = \text{número de veces en que ocurre la categoría } C_i, \quad i = 1, 2, \ldots, k,$$

entonces se dice que el vector aleatorio $(X_1, X_2, \ldots, X_k)$ tiene distribución multinomial de parámetros n, $p_1, p_2, \ldots, p_k$ y se le denota por $(X_1, X_2, \ldots, X_k) \sim Mul(n; p_1, p_2, \ldots, p_k)$. Detallamos seguidamente algunas de las propiedades de esta distribución.

Proposición 1.2. *Si* $(X_1, X_2, \ldots, X_k) \sim Mul(n; p_1, p_2, \ldots, p_k)$, *entonces:*

a) *La función de probabilidad (conjunta) de este vector viene dada por*

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \begin{cases} \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} & si\ (x_1, x_2, \ldots, x_k) \in R \\ 0 & en\ caso\ contrario \end{cases}$$

donde $R = \{(n_1, n_2, \ldots, n_k) \in \{0, 1, \ldots, n\}^k \ / \ \sum_{i=1}^k n_i = n\}$ *denota rango del vector*

b) $X_i \sim B(n, p),\ \forall i = 1, 2, \ldots, k$

c) $Cov(X_i, X_j) = -np_ip_j, \quad \forall i \neq j \in \{1, 2, \ldots, k\}$

<u>*Demostración:*</u> **a)** *La probabilidad de que en las primeras* x_1 *repeticiones ocurra* C_1, *en las siguientes* x_2 *repeticiones ocurra* C_2 *y así sucesivamente hasta que en las últimas* x_k *repeticiones ocurra* C_k *es por la independencia* $p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$. *Sin embargo, estas ocurrencias podrían darse de otras formas en términos del orden de ocurrencia de cada categoría. Todas las ordenaciones posibles de los n experimentos en donde* x_1 *serán de tipo* C_1, *y así sucesivamente hasta* x_k *del tipo* C_k, *vienen dadas por* $\frac{n!}{x_1! x_2! \ldots x_k!}$. *Por tanto, la probabilidad pedida viene dada por la fórmula en* **a)**.
b) *Basta notar que los experimentos que generan la multinomial podrían redefinirse como experimentos de Bernoulli. En efecto, si llamamos éxito a que ocurra la categoría* C_i *y fracaso a que ocurra cualquier otra categoría, el número de éxitos en las n repeticiones independientes tiene distribución binomial de parámetros n y* p_i. *Ella será entonces la distribución marginal de la v.a.* X_i.
c) *Basta notar que de juntar en una a las categorías* C_i *y* C_j, *con* $i \neq j$, *se tiene que*

$$X_i + X_j \sim B(n, p_i + p_j).$$

Así,

$$n(p_i + p_j)(1 - p_i - p_j) = V(X_i + X_j) = V(X_i) + V(X_j) + 2Cov(X_i, X_j)$$

$$np_i(1 - p_i) + np_j(1 - p_j) + 2Cov(X_i, X_j).$$

Un despeje directo en esta ecuación nos lleva a que $Cov(X_i, X_j) = -np_ip_j$. ■

Cabe comentar que las variables aleatorias δ_i definidas en (1.2), que denotan el número de veces en que la i-ésima unidad de la población física de tamaño N es seleccionada en una muestra al azar y con reemplazamiento de tamaño n, son todas v.a. con distribución

$B(n, \frac{1}{N})$. Más aún, si se tuviera interés en las frecuencias de selección de los elementos $i \neq j$ de la población, entonces no es difícil verificar que

$$(\delta_i, \delta_j, \delta_0) \sim Mul(n; \frac{1}{N}, \frac{1}{N}, 1 - \frac{2}{N}),$$

donde δ_0 denota la frecuencia de selecciones de otras unidades distintas a i y j. Note que estas v.a. no son independientes, desde que, por ejemplo:

$$P(\delta_j = y \mid \delta_i = x) = \frac{P(\delta_i = x, \delta_j = y, \delta_0 = n - x - y)}{P(\delta_i = x)} = C_y^{n-x}(\frac{1}{N-1})^y(1 - \frac{1}{N-1})^{n-y}$$

$$\neq C_y^n(\frac{1}{N})^y(1 - \frac{1}{N})^{n-y} = P(\delta_j = y),\ \forall x, y \in \{0, 1, \ldots, n\} \text{ con } x + y \leq n.$$

De manera general se cumple que

$$(\delta_1, \delta_2, \ldots, \delta_N) \sim Mul(n; \frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}).$$

1.3.3. La distribución hipergeométrica

Considere una población de N elementos, M de los cuales son de tipo A, y supongamos que se extrae al azar y sin reemplazamiento una muestra de n elementos de esta población. Si definimos

$$X = \text{ Número de elementos de tipo } A \text{ en la muestra,}$$

entonces se dice que X es una v.a. con distribución hipergeométrica de parámetros N, M y n y se le denota por $X \sim H(N, M, n)$.

Proposición 1.3. *Si $X \sim H(N, M, n)$, entonces:*

a) *La función de probabilidad de X viene dada por*

$$P_X(x) = P(X = x) = \begin{cases} \frac{C_x^M C_{n-x}^{N-M}}{C_n^N} & \text{si } x = 0, 1, 2, ..., n \\ 0 & \text{en otro caso,} \end{cases}$$

donde se conviene que $C_a^b = 0$, si $a > b$

b) $E(X) = n\frac{M}{N}$

c) $V(X) = n\frac{M}{N}(1 - \frac{M}{N})(\frac{N-n}{N-1})$

<u>Demostración:</u> **a)** *Supongamos, sin pérdida de generalidad, que $N - M < n < M$ (analice como ejercicio los otros casos) y consideremos un elemento cualesquiera x de su rango R_X, el cual por las condiciones dadas sería $R_X = \{0, 1, 2, \ldots, n\}$. Sea x un elemento cualquiera de R_X, entonces el evento $(X = x) = \{\omega \in \Omega \ / \ X(\omega) = x\}$, donde Ω denota espacio muestral conformado por todas las muestras o subconjuntos de n elementos que podríamos tomar de*

los N de la población, ocurre si y sólo si en la muestra x elementos poseen la característica A y $n-x$ no la poseen. Dado que cualquier subconjunto de tamaño n de la población tiene la misma probabilidad de ser seleccionado en la muestra, podríamos aplicar la definición clásica de probabilidad y escribir

$$P_X(x) = P(X = x) = \frac{n(X = x)}{n(\Omega)}.$$

Por tanto, $n(\Omega) = C_n^N$ y por el principio de multiplicación $n(X = x) = C_x^M C_{n-x}^{N-M}$ (pues, en la muestra, primero debemos seleccionar x de los M elementos que tienen la característica A y luego $n-x$ de los $N-M$ que tienen la característica A^c). Así, $P_X(x) = \frac{C_x^M C_{n-x}^{N-M}}{C_n^N}$.
b) *Mostraremos solo* **b)** *y dejaremos como ejercicio* **c)**, *el cual podría obtenerse con un procedimiento análogo. Como en la proposición anterior, asumiremos, sin pérdida de generalidad, que para $n \geq 2$ (si $n = 1$ el resultado es directo) se cumple que $N - M < n < M$. Dado $x \in R_X = \{0, 1, 2, \ldots, n\}$, el siguiente resultado directo de combinatorias nos será de utilidad:*

$$xC_x^M = x\frac{M!}{(M-x)!x(x-1)!} = \frac{M(M-1)!}{(M-x)!(x-1)!} = MC_{x-1}^{M-1}, \quad si\ x > 0.$$

Luego, al hacer en la sumatoria de abajo el cambio de variable $k = x - 1$, se tiene que

$$E(X) = \sum_{x=0}^{n} x\frac{C_x^M C_{n-x}^{N-M}}{C_n^N} = M\sum_{x=1}^{n} \frac{C_{x-1}^{M-1} C_{n-x}^{N-M}}{C_n^N}$$

$$= \frac{M}{C_n^N} C_{n-1}^{M-1} \sum_{k=0}^{n-1} \frac{C_k^{M-1} C_{n-1-k}^{N-M}}{C_{n-1}^{M-1}} = \frac{M}{C_n^N} C_{n-1}^{M-1} = n\frac{M}{N},$$

donde la última suma en la ecuación anterior es 1, pues se está sumando allí la función de probabilidad de una v.a. con distribución $H(N-1, M-1, n-1)$. ■

1.3.4. La distribución hipergeométrica multivariada

Esta es la extensión multivariada de la distribución anterior. Aquí, en lugar de estar la población de tamaño N dividida en dos clases (A y A^c), esta se particiona en k clases, a las que denotaremos por $C_1, C_2, \ldots, C_k$. Cada clase C_i posee M_i elementos, de tal manera que $N = M_1 + M_2 + \ldots + M_k$. Si seleccionamos ahora al azar y sin reemplazamiento n elementos de esta población y definimos las variables aleatorias

$$X_i = \text{número de elementos de la clase } C_i \text{ seleccionados en la muestra, } \ i = 1, 2, \ldots, k,$$

entonces se dice que el vector aleatorio $(X_1, X_2, \ldots, X_k)$ tiene distribución hipergeométrica multivariada de parámetros n, $M_1, M_2, \ldots, M_k$ y se le denota por $(X_1, X_2, \ldots, X_k) \sim Hmul(n; M_1, M_2, \ldots, M_k)$.

Proposición 1.4. *Si* $(X_1, X_2, \ldots, X_k) \sim Hmul(n; M_1, M_2, \ldots, M_k)$*, entonces:*

a) *La función de probabilidad (conjunta) de este vector viene dada por*

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{C_{x_1}^{M_1} C_{x_2}^{M_2} \ldots C_{x_k}^{M_k}}{C_n^N},$$

donde algunas de las combinatorias $C_a^b = 0$ *arriba son nulas si* $a > b$

b) $X_i \sim H(N, M_i, n), \quad \forall i = 1, 2, \ldots, k$

c) $Cov(X_i, X_j) = -\frac{nM_iM_j}{N^2}(\frac{N-n}{N-1}), \quad \forall i \neq j \in \{1, 2, \ldots, k\}$

d) *Si la muestra fuera tomada con reemplazamiento,*

$$(X_1, X_2, \ldots, X_k) \sim Mul(n; \frac{M_1}{N}, \frac{M_2}{N}, \ldots, \frac{M_k}{N})$$

Demostración: **a)** *El espacio muestral asociado a esta selección está constituido por todos los conjuntos de n elementos que se pueden formar con los N y, por tanto, tiene* C_n^N *elementos. El evento de interés tiene, por otro lado, en base al principio de multiplicación,* $C_{x_1}^{M_1} C_{x_2}^{M_2} \ldots C_{x_k}^{M_k}$ *elementos. Así, por la definición clásica, la probabilidad pedida es el cociente de estas cantidades.*
b) *Basta observar que de segmentarse la población en solo dos clases: la clase* A_i *de* M_i *elementos y la clase* A_i^c *de* $N - M_i$ *elementos, entonces* $X_i \sim H(N, M_i, n)$*.*
c) *Como en el multinomial, si juntamos juntamos dos categorías* A_i *y* A_j *(con* $i \neq j$*) en una sola, se tendrá que* $X_i + X_j \sim H(N, M_i + M_j, n)$*. Así,*

$$n(\frac{M_i + M_j}{N})(1 - \frac{M_i + M_j}{N})(\frac{N-n}{N-1}) = V(X_i + X_j)$$

$$= n(\frac{M_i}{N})(1 - \frac{M_i}{N})(\frac{N-n}{N-1}) + n(\frac{M_j}{N})(1 - \frac{M_j}{N})(\frac{N-n}{N-1}) + 2Cov(X_i, X_j).$$

Un despeje directo en esta ecuación nos lleva a que $Cov(X_i, X_j) = -\frac{nM_iM_j}{N^2}(\frac{N-n}{N-1})$*.*
d) *Si se admitiera reemplazamiento, cada selección generaría un experimento con k posibles resultados, siendo* $p_i = \frac{M_i}{N}$ *la probabilidad de que en el i-ésimo experimento se obtenga un elemento de la categoría* C_i*. Además, dada la independencia de estos experimentos por el reemplazo, el vector aleatorio* $(X_1, X_2, \ldots, X_k)$ *que cuenta las veces que ocurren cada una de estas k categorías en los n experimentos tendrá la distribución multinomial descrita.* ■

Por último, note que las v.a. δ_i discutidas en (1.2) tienen una naturaleza completamente distinta si la muestra se toma sin reemplazamiento. En efecto, si esta fuera la situación y se tuviera interés en la selección, por decir, de las unidades $i \neq j$ de la población física,

entonces para la distribución conjunta del vector $(\delta_i, \delta_j, \delta_0)$, que denota respectivamente a las frecuencias de selección de las unidades i, j u otras en la muestra, se cumpliría que

$$(\delta_i, \delta_j, \delta_0) \sim Hmul(n; 1, 1, N-2).$$

Aprecie que las v.a. δ_i y δ_j de este vector están ahora restringidas a tomar solo dos valores (0 o 1) y no son independientes desde que

$$P(\delta_j = 1 \mid \delta_i = 1) = \frac{P(\delta_i = 1, \delta_j = 1, \delta_0 = n-2)}{P(\delta_i = 1)} = \frac{n-1}{N-1} \neq \frac{n}{N} = P(\delta_j = 1),$$

ya que marginalmente $\delta_j \sim H(N, 1, n)$. En general, se cumplirá que

$$(\delta_1, \delta_2, \ldots, \delta_N) \sim HMul(n; 1, 1, \ldots, 1).$$

1.4. Esperanza, varianza y covarianza condicional

Discutiremos seguidamente una propiedad recurrente en varias aplicaciones del curso. Esta se refiere al cálculo indirecto de la media, varianza y covarianza mediante el condicionamiento de las variables de interés a un vector aleatorio $\mathbf{Z}$.

Proposición 1.5. *Si X, Y son dos v.a. con varianza finita y $\mathbf{Z}$ es un vector aleatorio, entonces:*

$$E(X) = E\left(E(X \mid \mathbf{Z})\right)$$

y

$$Cov(X, Y) = E\left(Cov(X, Y \mid \mathbf{Z})\right) + Cov\left(E(X \mid \mathbf{Z}), E(Y \mid \mathbf{Z})\right).$$

En particular,

$$V(X) = E\left(V(X \mid \mathbf{Z})\right) + V\left(E(X \mid \mathbf{Z})\right).$$

Demostración: Para probar la primera afirmación asumamos, sin pérdida de generalidad, que $\mathbf{Z}$ es un vector aleatorio discreto (que es el caso más recurrente en el muestreo). Entonces, sumando sobre todo valor posible del vector aleatorio $\mathbf{Z}$, se tiene que

$$E\left(E(X \mid \mathbf{Z})\right) = \sum_{\mathbf{z}} E(X \mid \mathbf{Z} = \mathbf{z}) P(\mathbf{Z} = \mathbf{z}) = \sum_{\mathbf{z}} \sum_{x \in R_X} x P(X = x \mid \mathbf{Z} = \mathbf{z}) P(\mathbf{Z} = \mathbf{z})$$

$$= \sum_{\mathbf{z}} \sum_{x \in R_X} x P(X = x, \mathbf{Z} = \mathbf{z}) = \sum_{x \in R_X} x \sum_{\mathbf{z}} P(X = x, \mathbf{Z} = \mathbf{z}) = \sum_{x \in R_X} x P(X = x) = E(X).$$

En relación con la covarianza, podríamos usar el resultado anterior y escribir

$$E\left(Cov(X, Y \mid \mathbf{Z})\right) = E\left(E(XY \mid \mathbf{Z}) - E(X \mid \mathbf{Z}) E(Y \mid \mathbf{Z})\right)$$

$$= E(XY) - E\left(E(X \mid \mathbf{Z}) E(Y \mid \mathbf{Z})\right)$$

$$Cov\left(E(X \mid \mathbf{Z}), E(Y \mid \mathbf{Z})\right) = E\left(E(X \mid \mathbf{Z})E(Y \mid \mathbf{Z})\right) - E\left(E(X \mid \mathbf{Z})\right)E\left(E(Y \mid \mathbf{Z})\right)$$
$$= E\left(E(X \mid \mathbf{Z})E(Y \mid \mathbf{Z})\right) - E(X)E(Y).$$

Sumándose ambos términos, el resultado es $E(XY) - E(X)E(Y)$, que no es sino la covarianza entre X e Y. ■

Ejemplo 1.1. *Un almacén contiene 6 cajas con la siguiente distribución:*

Caja (i)	1	2	3	4	5	6
Número de artículos (N_i)	60	43	97	80	120	100
Número de defectos (M_i)	5	4	6	5	15	10

Si para estimar la proporción de defectos en este almacén usted selecciona al azar una caja y extrae aleatoriamente y sin reemplazamiento un 20 % de sus artículos,

a) ¿Define tal procedimiento un estimador insesgado de la proporción buscada?

b) Obtenga la varianza del estimador propuesto.

Solución: a) Denotemos por δ_i a la variable indicadora que nos dice si la caja i ha sido ($\delta_i = 1$) o no ($\delta_i = 0$) seleccionada. Entonces, $(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6) \sim Mul(1; \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$. El estimador propuesto se puede escribir luego como $\hat{p} = \sum_{i=1}^{6} \bar{p}_i \delta_i$, donde $\bar{p}_i = \frac{X_i}{n_i}$ denota la proporción muestral de defectos en la caja i; $X_i \sim H(N_i, M_i, n_i)$ denota el número de defectos en la muestra de la caja i, y n_i es el tamaño de muestra para la caja i, el cual es igual al 20 % de N_i o a su entero superior, pensando como es racional que se desean minimizar costos.

Para el insesgamiento usaremos, tal como se ilustra seguidamente, la proposición 1.5:

$$E(\hat{p}) = E\left(E(\hat{p} \mid \delta_1, \delta_2, \ldots, \delta_6)\right) = E(\sum_{i=1}^{6} \frac{E(X_i)}{n_i}\delta_i) = \sum_{i=1}^{6} \frac{M_i}{N_i} E(\delta_i) = \frac{1}{6}\sum_{i=1}^{6} p_i,$$

siendo $p_i = \frac{M_i}{N_i}$ la proporción de defectos en la caja i. Dado que en general

$$E(\hat{p}) = \frac{1}{6}\sum_{i=1}^{6} \frac{M_i}{N_i} \neq \frac{\sum_{i=1}^{6} M_i}{\sum_{i=1}^{6} N_i} = p,$$

$\hat{p}$ no es un estimador insesgado de p.

Note que si conociéramos la distribución dada para este ejemplo (lo cual probablemente no ocurra y por eso se hace el muestreo), se tendría que $E(\hat{p}) = 0.0876$, valor que difiere de la verdadera proporción de defectos en el almacén que es de $p = 0.09$.

b) Por la proposición 1.5, $V(\hat{p}) = E\left(V(\hat{p} \mid \delta_1, \ldots, \delta_6)\right) + V\left(E(\hat{p} \mid \delta_1, \ldots, \delta_6)\right)$. Como las muestras en cada caja se pueden asumir independientes, se tendrá que

$$V(\hat{p} \mid \delta_1, \ldots, \delta_6) = \sum_{i=1}^{6} V(\bar{p}_i)\delta_i^2 = \sum_{i=1}^{6} \frac{1}{n_i}\frac{M_i}{N_i}(1 - \frac{M_i}{N_i})(\frac{N_i - n_i}{N_i - 1})\delta_i^2,$$

y así

$$E\left(V(\hat{p} \mid \delta_1, \ldots, \delta_6)\right) = \frac{1}{6}\sum_{i=1}^{6} \frac{1}{n_i}\frac{M_i}{N_i}(1 - \frac{M_i}{N_i})(\frac{N_i - n_i}{N_i - 1}) = \frac{1}{6}\sum_{i=1}^{6} \frac{1}{n_i}\frac{p_i(1-p_i)(N_i - n_i)}{n_i(N_i - 1)}.$$

Por otro lado, como $E(\hat{p} \mid \delta_1, \ldots, \delta_6) = \sum_{i=1}^{6} p_i\delta_i$, *se tiene que*

$$V\left(E(\hat{p} \mid \delta_1, \ldots, \delta_6)\right) = \sum_{i=1}^{6} p_i^2 V(\delta_i) + \sum_{\substack{i=1 \\ i\neq j}}^{6}\sum_{j=1}^{6} p_i p_j Cov(\delta_i, \delta_j) = \frac{5}{36}\sum_{i=1}^{6} p_i^2 - \frac{1}{36}\sum_{\substack{i=1 \\ i\neq j}}^{6}\sum_{j=1}^{6} p_i p_j.$$

Por tanto, la varianza pedida viene dada por

$$V(\hat{p}) = \frac{1}{6}\sum_{i=1}^{6} \frac{1}{n_i}\frac{p_i(1-p_i)(N_i - n_i)}{n_i(N_i - 1)} + \frac{5}{36}\sum_{i=1}^{6} p_i^2 - \frac{1}{36}\sum_{\substack{i=1 \\ j\neq i}}^{6}\sum_{j=1}^{6} p_i p_j.$$

En el caso que se conociera la distribución en el almacén, esta varianza podría evaluarse y vendría dada por $V(\hat{p}) = 0.004711247$. □

1.5. Selección de muestras al azar con y sin reemplazamiento

A lo largo del curso exploraremos diferentes librerías en R que nos permitirán no solo tomar muestras complejas, sino también analizarlas. En esta sección veremos las dos formas más básicas de seleccionar una muestra: las de tomarlas al azar con y sin reemplazamiento y en las que todos los elementos tendrán la misma probabilidad de selección. En adelante, a todo procedimiento de selección lo denominaremos un algoritmo de muestreo.

El algoritmo de muestreo con reemplazamiento es directo y se realiza utilizando la función de distribución empírica asociada a la selección de los elementos de una población física $\mathcal{P} = \{1, 2, \ldots, N\}$:

$$\hat{F}(i) = \frac{i}{N}.$$

Aquí basta generar de manera independiente n números aleatorios de una distribución uniforme en el intervalo $[0, 1]$, $u_1, u_2, \ldots, u_n$ y seleccionar las n unidades $i_1, i_2, \ldots, i_n$ en $\mathcal{P}$ mediante

$$i_k = \text{mín}\{i \in \mathcal{P} \ / \hat{F}(i) \geq u_k\}, \quad \forall k = 1, 2, \ldots, n.$$

En un muestreo sin reemplazamiento, el procedimiento anterior no es tan directo, pues la no restitución de los elementos previamente tomados modifica la función de distribución empírica asociada a la selección de los elementos de la población física, la cual se va también

modificando. Una manera de realizar esto es procediendo de forma secuencial; es decir, empezando por generar independientemente n números aleatorios $u_1, u_2, \ldots, u_n$ en el intervalo $[0, 1]$ y obteniendo, como antes, el primer elemento de la muestra mediante

$$i_1 = \text{mín}\{i \in \mathcal{P} \ / \hat{F}(i) \geq u_1\}.$$

Una vez seleccionado el k-ésimo elemento, i_k, el siguiente a incluir será

$$i_{k+1} = \text{mín}\{i \in \mathcal{P} \setminus \{i_1, i_2, \ldots, i_k\} \ / \frac{o(i)}{N-k} \geq u_{k+1}\}, \quad \forall k = 1, 2, \ldots, n-1,$$

donde $o(i)$ denota la posición ordinal que la unidad i ocupa en el conjunto $\mathcal{P} \setminus \{i_1, i_2, \ldots, i_k\}$. El proceso se repetirá luego hasta obtenerse i_n.

Otro algoritmo de muestreo sin reemplazamiento es el enumerativo. Este consiste en etiquetar cada una de las C_n^N muestras posibles, seleccionar al azar un número aleatorio $u \in [0, 1]$ y escoger la muestra cuya etiqueta k dividida entre N sea la primera en superar a u.

Como se aprecia, los procedimientos anteriores pueden resultar engorrosos, sobre todo si la muestra es sin reemplazamiento. Afortunadamente, se dispone en R del comando `sample`,el cual nos permite seleccionar muestras de manera directa. La sintaxis de este comando es

```
m = sample(x, size, replace, prob),
```

donde `x` denota un vector con los valores de alguna variable de interés que se evaluó en cada elemento de la población; `size` es el tamaño de muestra; `replace` es TRUE o FALSE, dependiendo si la muestra es con o sin reemplazamiento, respectivamente (argumento que por defecto es sin reemplazamiento), y `prob` es un vector con las probabilidades de selección para cada elemento en `x` (argumento opcional que por defecto asume que todos los elementos en la población tienen la misma probabilidad de selección). Si los valores de la variable en `x` no son de interés, sino que solo deseamos elegir n elementos de esta, el primer argumento de esta función puede también ser N, que es el tamaño de la población. Si escribimos en R

```
set.seed(12345)
(m = sample(80, 10))

##  [1] 58 70 60 69 35 13 25 38 53 71
```

`m` es un vector cuyas componentes corresponden a los elementos seleccionados en $\mathcal{P} = \{1, 2, \ldots, 80\}$ mediante un muestreo al azar y sin reemplazamiento de tamaño 10. Cabe aclarar que estamos fijando en la primera línea de comandos una semilla aleatoria. Esta será la misma semilla que usaremos, en lo posible, a lo largo del curso con el fin de garantizar que nuestros resultados sean replicables por parte del lector. Por otro lado, los paréntesis en la segunda línea de comandos indican que este resultado se mostrará en pantalla.

1.6. Ejercicios

1. Obtenga los intervalos de confianza de Wald y Wilson al 95 % para la proporción p de defectos de los artículos de una línea continua de producción, si al seleccionarse al azar 100 artículos de esta línea se encontraron 4 artículos defectuosos.

2. Juan, Pepe, Rosa, Luis y María participan en un sorteo donde se han de repartir entre ellos 4 vales de 50 soles cada uno.
a) Si Juan desea ganar algo, ¿qué le convendría más: un sorteo con o sin reemplazamiento?
b) Si la selección se hace con reemplazamiento, ¿qué probabilidad hay de que Juan gane 1 vale y Rosa 2? ¿Es esta probabilidad la misma a que Juan gane los cuatro vales?
c) Bajo reemplazamiento, ¿con qué probabilidad solo Rosa y Luis ganarán vales?
d) Halle, en el caso de que el sorteo se haga con reemplazamiento, el monto que esperará obtener Juan en el sorteo.

3. a) Sea X una variable aleatoria con distribución binomial de parámetros N y p, y supongamos que la distribución condicional de una v.a. Y, dado que $X = x$, es hipergeométrica con $Y|_{X=x} \sim H(N, x, n)$. Demuestre que $Y \sim B(n, p)$.
b) Suponga que en un estudio sobre la prevalencia de una enfermedad (proporción p de personas que la padecen) se piensa tomar una muestra al azar con reemplazamiento de tamaño 420. Un estadístico opina que esto es excesivo, pues conocer si las personas tienen o no la enfermedad implicará aplicar una prueba cara y de logística algo complicada. Dado que ya se han enviado cartas a las personas seleccionadas, el estadístico suguiere tomar más bien un muestreo al azar y sin reemplazamiento de tamaño 80 de la población inicialmente contactada. Si se acepta la sugerencia del estadístico y si p es 0.1, ¿con qué probabilidad se encontrará en la muestra más de 5 personas que padezcan la enfermedad?

4. La producción diaria de una fábrica, que es de 200 artículos, contiene 12 artículos con un defecto de tipo A y 8 artículos con un defecto de tipo B. Si usted adquiere al azar y sin reemplazamiento 20 de estos artículos y sabe que cada artículo bueno le reportará una utilidad de 25 soles; mientras que cada artículo con defectos de tipo A y B le reportará una pérdida de 5 y 10 soles, respectivamente,
a) ¿Con qué probabilidad obtendrá una utilidad de 400 soles al vender los 20 artículos?
b) Halle el valor esperado y la desviación estándar de la utilidad de venta de los 20 artículos.

5. En un experimento se colocan, uno a uno, 20 ratones en una caja con 8 puertas idénticas. Dos de las puertas conducen a un premio; una a un castigo, y las otras son neutras. Sean X_P, X_C y X_N el número de estos ratones que eligen la puerta con premio, castigo y neutra, respectivamente, en su primer intento.
a) ¿Cuál es la distribución conjunta de estas variables aleatorias?
b) Halle e interprete la correlación de Pearson entre X_P y X_N.

6. Dos encuestadoras han seleccionado al azar y sin reemplazamiento muestras de tamaños 20 y 10 en una población de 50 personas. Halle la función de probabilidad, valor esperado y varianza del número de personas que serán entrevistadas por ambas encuestadoras.

7. Un encuestador tiene asignado un área de trabajo de 100 viviendas, donde se sabe que el 10 % de estas presentan una característica A que solo se podrá conocer durante la entrevista. El entrevistador visitará casa por casa y aplicará una encuesta más larga a las viviendas que poseen la característica A. Suponga que el encuestador tiene una cuota de 5 viviendas con la característica A, luego de lo cual será reemplazado por otro encuestador.
a) Halle la función de probabilidad del número de entrevistas que realizará el entrevistador.
b) Suponga que el tiempo en minutos que emplea el entrevistador en realizar una encuesta a una vivienda, sin y con la característica de interés, es una v.a. con distribución normal de media 8 minutos y desviación estándar de 2 minutos y media 15 minutos y desviación estándar de 4 minutos, respectivamente. Halle el tiempo efectivo que se espera le tome al entrevistador realizar todas sus encuestas.

8. Suponga que 4 cápsulas de un medicamento genérico fueron mezcladas con 20 de marca y luego distribuidas al azar en 4 cajas de 6 cápsulas cada una. Una manera de pensar la distribución de las cápsulas en las cajas es secuencialmente; esto es, eligiendo primero al azar y sin reemplazamiento 6 cápsulas para colocarlas en una caja, que etiquetaremos 1; seleccionando luego al azar y sin reemplazamiento otras 6 cápsulas de las 18 restantes para colocarlas en una caja 2, y seleccionando finalmente al azar y sin reemplazamiento 6 de las 12 cápsulas restantes para colocarlas en una caja 3. Las cápsulas sobrantes conformarán la caja 4.
a) Halle la función de probabilidad del número de cápsulas del medicamento genérico que contendrá la caja etiquetada como 1.
b) Halle la probabilidad de que solamente la caja 3 contenga cápsulas del medicamento genérico. ¿Es esta probabilidad la misma si se tratara de la caja 1?
c) Verifique que la probabilidad de que una caja contenga `x` cápsulas genéricas es siempre la misma al margen de la etiqueta que tenga la caja. Ello puede hacerlo calculando esta probabilidad para cada etiquetado y cada valor posible `x`. Como ayuda, puede usar la función `dhyper(x,M,N-M,n)` de R que le permite hallar la probabilidad de que una v.a. $X \sim H(N, M, n)$ tome el valor `x`.
d) Muestre que la función de probabilidad conjunta del número de cápsulas del medicamento genérico que contendrá cada una de las 4 cajas (X_1, X_2, X_3, X_4) viene dada por

$$P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4) = \frac{C^6_{x_1} C^6_{x_2} C^6_{x_3} C^6_{x_4}}{C^{24}_4}.$$

e) Halle la función de probabilidad del número de cajas que contendrá alguna cápsula del medicamento genérico.

9. Un peaje tiene 10 casetas de pago, 3 de las cuales son exclusivas para buses y camiones y el resto se destina a solo autos. Suponga que el 20 % de los vehículos que pasan por este peaje son buses o camiones y el resto autos, y que los vehículos tienen igual probabilidad de escoger cualquier caseta que les corresponda. Si la administradora le informa que en un día pasaron por el peaje 800 vehículos, que en las 3 casetas para buses y camiones se registraron 56, 65 y 60 unidades, pero no le informa sobre la distribución del número de autos en las demás casetas,
a) ¿Cuál será la distribución conjunta del número de autos que pasaron durante ese día por las casetas correspondientes?
b) ¿Que tán probable fué que se hallan obtenido estos resultados para las casetas de buses o camiones?
c) ¿Cuántos autos se espera que hayan pasado por el peaje 4 (de autos) durante ese día?

10. Sea $(X_1, X_2, \ldots, X_k) \sim HMul(n; M_1, M_2, \ldots, M_k)$ y supongamos seleccionamos tres componentes cualesquieras i, j y m de este vector. ¿Es cierto que el vector aleatorio (X_i, X_j, X_m) tiene también una distribución hipergeométrica multivariada?

11. Se ha creado una nueva agrupación política producto de la fusión de 3 partidos políticos A, B y C. En esta nueva agrupación participan 10 personas del partido A, 20 personas del partido B y 14 personas del partido C. Para crear esta nueva agrupación cada miembro de estos partidos aportó, respectivamente, 100, 500 y 300 u.m. Si usted selecciona al azar y sin reemplazamiento a 10 personas de esta nueva agrupación para aplicarles una encuesta,
a) ¿Con qué probabilidad la mitad de los encuestados serán ex integrantes del partido C?
b) ¿Cuál es la probabilidad de que todas las personas encuestadas menos una hayan sido integrantes del partido B?
c) ¿Cuál es el monto total de aporte que se esperará reporten las personas en la encuesta?
d) Si le informan, luego de tomarse la muestra, que solo 3 personas que pertenecieron al partido A fueron encuestadas, ¿cuántas personas del otrora partido B se espera hayan sido encuestadas?
e) Suponga que el 80 %, 30 % y 50 % de las personas de los otrora partidos A, B y C tenía interés en formar parte del Ejecutivo.

e.1) ¿Cuántos encuestados, ex-integrantes del partido C, se esperan tengan interés en el Ejecutivo?

e.2) En general, ¿cuántas de las personas seleccionados para la encuesta espera que tengan interés en el Ejecutivo?

e.3) Si se propone como estimador de la proporción de interés en el Ejecutivo a la correspondiente proporción muestral en la encuesta, ¿forma este un estimador insesgado?

e.4) ¿Cuál es la varianza del estimador propuesto en e.3)?

12. Proponga, para el ejemplo 1.1, un estimador insesgado de la proporción buscada y calcule su varianza.

13. En este ejercicio, tomado de Valdivieso (2017), una empresa recibe lotes de 500 artículos de un fabricante y utiliza el siguiente plan de muestreo doble para la inspección de recibo:

i) Se toma una muestra al azar y sin reemplazamiento de 15 unidades. Si ningún artículo es defectuoso, se acepta el lote; si se encuentran 3 o más artículos defectuosos, se lo rechaza; en cualquier otro caso se toma una segunda muestra de 13 unidades.

ii) Si el número total de unidades defectuosas (en ambas muestras) es mayor que 3, se rechaza el lote.

iii) Finalmente, si se rechaza el lote, se inspeccionan el 100 % de sus unidades y el fabricante debe cambiar las unidades defectuosas por buenas y pagar los costos de inspección.

Si los lotes recibidos tienen un 5 % de unidades defectuosas y el costo de inspección de una unidad es de un sol, halle:

a) La probabilidad de rechazar el lote.

b) El gasto esperado por inspección por parte de la empresa y del fabricante.

14. Un congreso cuenta con la participación de N instituciones, siendo M_i el número de participantes de la i-ésima institución. A fin de recabar información de los participantes y sobre todo de sus instituciones, se ha diseñado una encuesta por muestreo en la que se seleccionarán al azar a n personas, pero en la que solo se preguntará sobre la institución a la primera persona que se encuentre de cada institución. Sea Ne el número de instituciones distintas que se encuentran en la muestra.
a) ¿Con qué probabilidad la muestra estará conformada por solo participantes de las tres primeras instituciones?
b) Si solo la primera institución tiene n o más participantes, ¿con qué probabilidad $Ne = 1$?
c) ¿Cuántas instituciones se esperará encuestar? SUG: Considere las v.a. indicadoras $1_{\{X_i>0\}}$, donde X_i denota el número de personas de la institución i que serán encuestadas.
d) Si la distribución de los participantes en el congreso fue la siguiente:

Institución (i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Número de participantes (M_i)	17	8	3	4	6	9	12	14	1	2	1	4	2	10	2	5

y la muestra es de tamaño $n = 16$, obtenga las cantidades anteriormente pedidas y calcule la probabilidad de que $Ne = 2$.

15. Con el fin de obtener una muestra al azar y sin reemplazamiento que corresponda al 20 % de una población de tamaño $N = 100$, un alumno ha desarrollado el siguiente algoritmo: simular 100 números aleatorios en el intervalo $[0, 1]$ y tomar como muestra aquellas unidades $i \in \mathcal{P} = \{1, 2, \ldots, 100\}$, tales que sus correspondientes números aleatorios sean menores o iguales a 0.2. ¿Es correcto este algoritmo de muestreo? Justifique.

16. Un gran complejo turístico tiene 12 parques temáticos que se pueden visitar uno por día. Un turista solo tiene 4 días de estadía, por lo cual elige al azar 4 de estos parques. Tiempo después de comprar sus entradas se entera de que 3 de los parques cobran parqueo.
a) Halle la función de probabilidad del número de parqueos que tendrá que pagar el turista y calcule su valor esperado.
b) Simule el número de parqueos que tendrá que pagar el turista durante su estadía.
c) Si se propone el siguiente algoritmo de muestreo:

```
m <- u <- runif(4)
for(i in 1:4){m[i] = min(which((1:12/12>u[i])==TRUE))}
sum(as.numeric(m<=9))
```

¿Es este algoritmo de muestreo adecuado para lo que se pide en b)?

17. Si selecciona al azar 6 compañeros de su salón de clase basándose en la lista de alumnos del salón,
a) ¿Con qué probabilidad usted saldrá elegido de tomarse la muestra sin reemplazamiento?
b) ¿Con qué probabilidad algunos de sus compañeros serán elegidos en más de una oportunidad de tomarse la muestra con reemplazamiento?

Capítulo 2

Muestreo aleatorio simple

En un muestreo aleatorio simple (MAS) toda muestra de tamaño n tiene la misma probabilidad de ser seleccionada, lo cual corresponde teóricamente a la noción de muestra aleatoria dada en la sección anterior si la población es infinita. En la práctica las poblaciones son finitas, digamos con N elementos. Aquí veremos cómo tomar en cuenta este hecho y nos interesará encontrar tamaños de muestra y errores de estimación para tres de los parámetros más frecuentemente referidos en un estudio inferencial, la media poblacional μ, el total poblacional τ y la proporción de elementos p de la población que comparten alguna característica particular. Para ser más precisos enfatizaremos sobre todo el primero y último de estos parámetros, pues el análisis para el total poblacional

$$\tau = N\mu \quad \text{o} \quad \tau = Np$$

es directamente deducible de los de μ y p.

2.1. Muestreo con y sin reemplazamiento

Existen dos esquemas de muestreo aleatorio simple importantes: el muestreo aleatorio simple con reemplazamiento, que lo denotaremos en adelante por MASc, y el muestreo aleatorio simple sin reemplazamiento, que lo denotaremos en adelante por MASs. En la sección 1.5 adelantamos ya varias de las características de estos esquemas, así como algunos de sus algoritmos de muestreo; es decir, cómo realizar el muestreo en la práctica. En esta sección nos enfocaremos más en el análisis de las unidades seleccionadas cuando en ellas se desee estudiar una o más variables de interés.

Con base en un enfoque basado en el diseño, consideremos primero para ello una población física $\mathcal{P} = \{1, 2, \ldots, N\}$ de tamaño N a cuyos elementos los estamos identificando, por simplicidad, con los números naturales positivos. A estos que pudieran ser sujetos, eventos, materiales, escuelas, países, etc, los llamaremos unidades. Sobre estas unidades mediremos

una variable estadística y para generar la población estadística $\mathcal{P}_y$ constituida por todos los valores de y en $\mathcal{P}$; es decir,

$$\mathcal{P}_y = \{y_1, y_2, \ldots, y_N\},$$

siendo y_i el valor de y para la unidad i. Note que algunos de estos valores pueden repetirse, lo cual no ocurre en $\mathcal{P}$. Sea $n < N$ el tamaño de muestra a seleccionarse.

En un esquema MASc, las unidades se seleccionan al azar una a una de la población, con la peculiaridad de que estos son repuestos o reemplazados en cada etapa de selección. Así, una unidad cualesquiera $j \in \mathcal{P}$ podría ser elegida en más de una oportunidad. Por otro lado, en el esquema MASs, las unidades seleccionadas no se reponen y, por tanto, una unidad cualesquiera $j \in \mathcal{P}$ podría ser elegida en a lo más una oportunidad. En este caso note que seleccionar las unidades una a una hasta completar la muestra equivale a seleccionar toda la muestra de una sola vez. La ventaja del diseño MASc es que las variables aleatorias definidas en (1.1) y asociadas a los valores de y en las unidades seleccionados son variables independientes. En efecto, esto se sigue desde que para cualquier par de selecciones $j < k$ y cualquier par de elementos $y_p, y_q \in \mathcal{P}_y$ de la población estadística:

$$P(Y_j = y_p, Y_k = y_q) = P(Y_k = y_q \mid Y_j = y_p)P(Y_j = y_p) = P(Y_k = y_q)P(Y_j = y_p).$$

En un MASs, por otro lado, lo anterior no siempre se cumple, ya que, por ejemplo,

$$P(Y_2 = y_q \mid Y_1 = y_p) = \frac{1}{N-1} \neq \frac{1}{N} = P(Y_2 = y_q)$$

en el que caso de que los elementos de la población estadística sean todos distintos.

Si bien la falta de independencia en un MASs puede acarrear problemas técnicos, este es en la práctica el esquema más utilizado porque garantiza siempre selecciones distintas en $\mathcal{P}$.

Enfaticemos ahora el estudio y las propiedades de dos de los estimadores más recurrentes en el muestreo, la media y la varianza muestrales

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{N} y_i\delta_i \quad \text{y} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2\delta_i,$$

donde recordemos que δ_i es una variable aleatoria que cuenta el número de veces que la unidad i de $\mathcal{P}$ es seleccionada en la muestra.

Tanto en el MASc como en el MASs, estas estadísticas constituyen los estimadores naturales de la media poblacional

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} y_i$$

y varianza poblacional

$$\sigma_N^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_N)^2 \quad \text{o} \quad \sigma_{N-1}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_N)^2.$$

En adelante, para una mejor comprensión, convendremos en denotar las variables aleatorias con letras mayúsculas (a excepción de los δ_i) y con letras minúsculas las no aleatorias.

Antes de analizar algunas propiedades de los estimadores $\bar{Y}$ y S^2, es útil recordar que el vector aleatorio de frecuencias de conteo por unidad de la muestra $(\delta_1, \delta_2, \ldots, \delta_N)$ tiene una distribución multinomial o hipergeométrica multivariada, dependiendo de si el esquema es un MASc o un MASs, respectivamente. Más aún, por lo visto en (1.2), tanto la media como la varianza muestral podrían escribirse alternativamente como

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

y

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

donde $Y_1, Y_2, \ldots, Y_n$ denotan los valores que secuencialmente la variable estadística en estudio y podría tomar en cada selección de la muestra. La proposición siguiente nos brinda algunas propiedades de estas últimas variables aleatorias.

Proposición 2.1.

a) *En un MASc, las v.a. $Y_1, Y_2, \ldots, Y_n$ son independientes e idénticamente distribuidas con media $E(Y_i) = \mu_N$ y varianza $V(Y_i) = \sigma_N^2$.*

b) *En un MASs, las v.a. $Y_1, Y_2, \ldots, Y_n$ son idénticamente distribuidas con media $E(Y_i) = \mu_N$, varianza $V(Y_i) = \sigma_N^2$ y se cumple que $Cov(Y_i, Y_j) = -\frac{1}{N}\sigma_{N-1}^2$, $\forall i \neq j$.*

Demostración: Supongamos, sin pérdida de generalidad, que todos los elementos en $\mathcal{P}_y$ son distintos.

a) *La independencia ya se analizó. Que las v.a. $Y_1, Y_2, \ldots, Y_n$ tengan la misma distribución de media μ_N y varianza σ_N^2 es, por otro lado, consecuencia directa de que la distribución de cualesquiera de estas variables, digamos Y_i, viene definida por la función de probabilidad*

$$P_{Y_i}(y) = P(Y_i = y) = \begin{cases} \frac{1}{N} & si\ y = y_1, y_2, y_3, \ldots, y_N \\ 0 & en\ otro\ caso \end{cases} \tag{2.1}$$

b) *Claramente, como la selección es secuencial, Y_1 tiene la distribución (2.1). Más aún, condicionando y trabajando inductivamente, se puede probar que la distribución de cualesquiera de las variables $Y_1, Y_2, \ldots, Y_n$, digamos Y_i, tiene la función de probabilidad dada en (2.1). Como podemos ver, para cualquier $j \in \mathcal{P}$:*

$$P(Y_2 = y_j) = \sum_{i=1}^{N} P(Y_2 = y_j \mid Y_1 = y_i)P(Y_1 = y_i)$$

$$= \sum_{\substack{i=1 \\ i\neq j}}^{N} P(Y_2 = y_j \mid Y_1 = y_i)\frac{1}{N} = \sum_{\substack{i=1 \\ i\neq j}}^{N} \frac{1}{N-1}\frac{1}{N} = \frac{1}{N}.$$

Otra manera de ver lo anterior y que nos servirá también para las otras afirmaciones es notando que la distribución conjunta del vector $(Y_1, Y_2, \ldots, Y_n)$ *viene dada por*

$$P(Y_1 = y_{j1}, Y_2 = y_{j2}, \ldots, Y_n = y_{jn})$$

$$= P(Y_n = y_{jn} \mid Y_1 = y_{j1}, \ldots, Y_{n-1} = y_{j(n-1)}) \ldots P(Y_2 = y_{j2} \mid Y_1 = y_{j1})P(Y_1 = y_{j1})$$

$$= \frac{1}{N-n+1} \times \frac{1}{N-n+2} \times \ldots \times \frac{1}{N-1} \times \frac{1}{N},$$

cualesquiera sea $k \in \{1, 2 \ldots, n\}$ *e* $y_{jk} \in \mathcal{P}_y$. *De esta distribución conjunta se pueden hallar distintas marginales, como la de la v.a.* Y_i, *la cual se obtiene sumando la última función de probabilidad conjunta sobre todos los valores de las demás variables. Estas sumas contienen* $(N-1)(N-2)\ldots(N-n+1)$ *términos, por lo cual su resultado nos dará* $\frac{1}{N}$, *que es precisamente la misma distribución que en el caso MASc. Por tal razón, las* Y_i*'s tienen la misma media y varianzas anteriores. Podemos también, por otro lado, hallar la distribución conjunta del vector* (Y_i, Y_j) *con* $i \neq j$. *Esta viene dada por la suma de la distribución conjunta sobre todos los valores de las demás* $n-2$ *variables que no contengan los valores donde se evalúan* Y_i *e* Y_j. *Estas sumas, como no es difícil ver, contienen* $(N-2)(N-3)\ldots(N-n+1)$ *términos, de aquí que se tenga que*

$$P(Y_i = y_p, Y_j = y_q) = \frac{(N-2)(N-3)\ldots(N-n+1)}{(N-n+1)(N-n+2)\ldots(N-1)N} = \frac{1}{N(N-1)}, \forall p \neq q \in \mathcal{P}.$$

Consecuentemente,

$$Cov(Y_i, Y_j) = E((Y_i - \mu_N)(Y_j - \mu_N)) = \sum_{p=1}^{N}\sum_{q=1}^{N}(y_p - \mu_N)(y_q - \mu_N)P(Y_i = y_p, Y_j = y_q)$$

$$= \sum_{\substack{p=1 \\ q\neq p}}^{N}\sum_{q=1}^{N}(y_p - \mu_N)(y_q - \mu_N)\frac{1}{N(N-1)} = \frac{1}{N(N-1)}\sum_{p=1}^{N}(y_p - \mu_N)(\sum_{q=1}^{N}(y_q - \mu_N) - (y_p - \mu_N))$$

$$= \tfrac{1}{N(N-1)}((\textstyle\sum_{p=1}^{N}(y_p - \mu_N))^2 - \sum_{p=1}^{N}(y_p - \mu_N)^2) = -\tfrac{1}{N}\sigma^2_{N-1}.$$ ■

Ejemplo 2.1. *Considere una población de sujetos* $\mathcal{P} = \{1, 2, 3, 4, 5, 6, 7\}$ *y la población estadística* $\mathcal{P}_y = \{12, 32, 18, 37, 22, 18, 28\}$ *asociada a la edad y de ellos en años. Suponga ahora que se toma un MAS con* $n = 3$. *Halle la distribución muestral de la media y varianza para esta muestra y verifique que estos son estimadores insesgados. Realice esto para los dos esquemas de muestreo estudiados.*

Solución: *La media y varianza poblacionales de y vienen dadas por* $\mu_7 = 23.9$, $\sigma_6^2 = 78.1$ *y* $\sigma_7^2 = 67$. *En un MASc tenemos, de tomarse en cuenta el orden, un total de* $7^3 = 343$ *muestras posibles; mientras que en un MASs tenemos un total de* $C_3^7 = 35$. *Nosotros desarrollaremos aquí el caso de un MASs y el otro esquema quedará como ejercicio para el lector. Como ayuda utilizaremos el paquete* `combinat` *de R. Dado que en este problema precisamos obtener la distribución muestral de la media y varianza muestrales, apelaremos al uso del comando* `combn` *y obtendremos para cada posible muestra tanto su media, varianza y probabilidad de selección. El código respectivo se muestra seguidamente y los resultados se resumen en los cuadros 2.1, 2.2 y 2.3.*

```
library(combinat)
options(digits=3)
ypop = c(12, 32, 18, 37, 22, 18, 28)
samplesMASs = t(as.matrix(combn(ypop,3)))
ybar = apply(samplesMASs,1,mean)
s2 = apply(samplesMASs,1,var)
probs = rep(1/length(ybar), length(ybar))
bsamplesMASs = cbind(samplesMASs,ybar,s2,probs)
pp1 = aggregate(bsamplesMASs[,6],by = list(bsamplesMASs[,4]),sum)
colnames(pp1) = c("Media muestral","Probabilidad")
pp2 = aggregate(bsamplesMASs[,6],by = list(bsamplesMASs[,5]),sum)
colnames(pp2) = c("Varianza muestral","Probabilidad")
```

Cabe comentar que si la muestra fuese con reemplazamiento, podríamos encontrar los índices de todas las posibles muestras con el comando `expand.grid(rep(list(1:7),3))`.
Según las tablas mostradas, los valores esperados de la media y varianza muestrales vendrán dados, respectivamente, por

```
c(sum(pp1[,1]*pp1[,2]),sum(pp2[,1]*pp2[,2]))

## [1] 23.9 78.1
```

mientras que la varianza de la media muestral es

```
sum(((pp1[,1] - sum(pp1[,1]*pp1[,2]))^2)*pp1[,2])

## [1] 14.9
```

Esto nos indica que la media muestral $\bar{Y}$ *es efectivamente un estimador insesgado de* μ_7; *mientras que la varianza muestral* S^2 *es un estimador insesgado de* σ_6^2.

□

Muestra				Mediam	Varm	Probs	Muestra				Mediam	Varm	Probs
1	12	32	18	20.7	105.3	0.0286	19	32	18	28	26	52	0.0286
2	12	32	37	27	175	0.0286	20	32	37	22	30.3	58.3	0.0286
3	12	32	22	22	100	0.0286	21	32	37	18	29	97	0.0286
4	12	32	18	20.7	105.3	0.0286	22	32	37	28	32.3	20.3	0.0286
5	12	32	28	24	112	0.0286	23	32	22	18	24	52	0.0286
6	12	18	37	22.3	170.3	0.0286	24	32	22	28	27.3	25.3	0.0286
7	12	18	22	17.3	25.3	0.0286	25	32	18	28	26	52	0.0286
8	12	18	18	16	12	0.0286	26	18	37	22	25.7	100.3	0.0286
9	12	18	28	19.3	65.3	0.0286	27	18	37	18	24.3	120.3	0.0286
10	12	37	22	23.7	158.3	0.0286	28	18	37	28	27.7	90.3	0.0286
11	12	37	18	22.3	170.3	0.0286	29	18	22	18	19.3	5.3	0.0286
12	12	37	28	25.7	160.3	0.0286	30	18	22	28	22.7	25.3	0.0286
13	12	22	18	17.3	25.3	0.0286	31	18	18	28	21.3	33.3	0.0286
14	12	22	28	20.7	65.3	0.0286	32	37	22	18	25.7	100.3	0.0286
15	12	18	28	19.3	65.3	0.0286	33	37	22	28	29	57	0.0286
16	32	18	37	29	97	0.0286	34	37	18	28	27.7	90.3	0.0286
17	32	18	22	24	52	0.0286	35	22	18	28	22.7	25.3	0.0286
18	32	18	18	22.7	65.3	0.0286							

Cuadro 2.1: Probabilidades, medias y varianzas de todas las posibles muestras en un MASs para el ejemplo 2.1

	Media muestral	Probabilidad
1	16.000	0.029
2	17.333	0.057
3	19.333	0.086
4	20.667	0.086
5	21.333	0.029
6	22.000	0.029
7	22.333	0.057
8	22.667	0.086
9	23.667	0.029
10	24.000	0.086
11	24.333	0.029
12	25.667	0.086
13	26.000	0.057
14	27.000	0.029
15	27.333	0.029
16	27.667	0.057
17	29.000	0.086
18	30.333	0.029
19	32.333	0.029

Cuadro 2.2: Distribución de la media muestral para el ejemplo 2.1

	Varianza muestral	Probabilidad
1	5.333	0.029
2	12.000	0.029
3	20.333	0.029
4	25.333	0.143
5	33.333	0.029
6	52.000	0.114
7	57.000	0.029
8	58.333	0.029
9	65.333	0.114
10	90.333	0.057
11	97.000	0.057
12	100.000	0.029
13	100.333	0.057
14	105.333	0.057
15	112.000	0.029
16	120.333	0.029
17	158.333	0.029
18	160.333	0.029
19	170.333	0.057
20	175.000	0.029

Cuadro 2.3: Distribución de la varianza muestral para el ejemplo 2.1

Como el ejemplo anterior lo sugiere, tenemos las siguientes propiedades en un MAS.

Proposición 2.2. *La media muestral $\bar{Y}$ es un estimador insesgado de la media poblacional μ_N y se tiene que*

a) $V(\bar{Y}) = \frac{\sigma_N^2}{n}$ *en un MASc*

b) $V(\bar{Y}) = (1 - \frac{n}{N})\frac{\sigma_{N-1}^2}{n}$ *en un MASs*

La demostración de la proposición anterior es directa y puede deducirse de la demostración del siguiente resultado de suma importancia.

Proposición 2.3.

a) *La media muestral es el MELI (mejor estimador lineal e insesgado) de la media poblacional.*

b) *La varianza muestral es un estimador insesgado de σ_N^2 para un MASc y de σ_{N-1}^2 para un MASs.*

Demostración: Puesto que la demostración de esta proposición es directa en el caso MASc, la dejaremos como ejercicio. Nosotros centraremos nuestra atención en el caso MASs.
a) *Sea $\hat{\mu}_N$ un estimador lineal arbitrario de la media poblacional; es decir, un estimador de la forma $\hat{\mu}_N = \sum_{i=1}^n c_i Y_i$, donde las constantes c_i que la definen son arbitrarias. Para que este sea un estimador insesgado se debe satisfacer*

$$\mu_N = E(\hat{\mu}_N) = E(\sum_{i=1}^n c_i Y_i) = \sum_{i=1}^n c_i E(Y_i) = \mu_N \sum_{i=1}^n c_i;$$

es decir, las constantes c_i deben sumar 1. Por otro lado, la varianza de este estimador lineal viene dado por

$$V(\hat{\mu}_N) = \sum_{i=1}^n c_i^2 V(Y_i) + \sum_{\substack{i=1 \\ j \neq i}}^n \sum_{j=1}^n c_i c_j Cov(Y_i, Y_j)$$

o más explícitamente, de lo visto en la proposición 2.1, por

$$V(\hat{\mu}_N) = \sigma_N^2 \sum_{i=1}^n c_i^2 - \frac{1}{N}\sigma_{N-1}^2 \sum_{\substack{i=1 \\ j \neq i}}^n \sum_{j=1}^n c_i c_j = \frac{N-1}{N}\sigma_{N-1}^2 \sum_{i=1}^n c_i^2 - \frac{1}{N}\sigma_{N-1}^2 (\sum_{i=1}^n \sum_{j=1}^n c_i c_j - \sum_{i=1}^n c_i^2)$$

$$= \sigma_{N-1}^2 (\sum_{i=1}^n c_i^2 - \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^n c_i c_j). \quad (*)$$

Por tanto, el MELI de μ_N se obtendrá al hallar las constantes c_i que resuelvan el siguiente problema de optimización:

$$\min_{s.a\sum_{i=1}^n c_i=1} \sum_{i=1}^n c_i^2 - \frac{1}{N}\sum_{i=1}^n\sum_{j=1}^n c_i c_j.$$

Dada la convexidad de la función objetivo, bastará considerar las condiciones de primer orden del lagrangiano de esta función, el cual viene dado por

$$l = \sum_{i=1}^n c_i^2 - \frac{1}{N}\sum_{i=1}^n\sum_{j=1}^n c_i c_j + \lambda(1 - \sum_{i=1}^n c_i).$$

De las derivadas parciales con respecto a c_k se obtiene que

$$0 = \frac{\partial l}{\partial c_k} = 2c_k - \frac{2}{N}\sum_{i=1}^n c_i - \lambda,$$

de donde $c_k = \frac{1}{N} + \frac{\lambda}{2}$. De la condición de insesgamiento, el multiplicador de Lagrange óptimo resulta ser $\lambda = \frac{2}{n}(1 - \frac{n}{N})$, el cual al reemplazarlo en la expresión previa nos da

$$c_k = \frac{1}{N} + \frac{1}{n}(1 - \frac{n}{N}) = \frac{1}{n}.$$

Consecuentemente, el MELI de μ_N es $\bar{Y}$. Más aún, la varianza de este estimador es por $(*)$

$$V(\bar{Y}) = (1 - \frac{n}{N})\frac{\sigma_{N-1}^2}{n}.$$

b) *Puesto que $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n-1}(\sum_{i=1}^n Y_i^2 - n\bar{Y}^2)$, se tiene que en un MASs*

$$E(S^2) = \frac{1}{n-1}(\sum_{i=1}^n E(Y_i^2) - nE(\bar{Y}^2))$$

$$= \frac{1}{n-1}(\sum_{i=1}^n (V(Y_i) + E(Y_i)^2) - n(V(\bar{Y}) + E(\bar{Y})^2))$$

$$= \frac{1}{n-1}(\sum_{i=1}^n (\sigma_N^2 + \mu_N^2) - n((1 - \frac{n}{N})\frac{\sigma_{N-1}^2}{n} + \mu_N^2))$$

$$= \frac{1}{n-1}(n(\frac{N-1}{N}\sigma_{N-1}^2 + \mu_N^2) - n((1 - \frac{n}{N})\frac{\sigma_{N-1}^2}{n} + \mu_N^2)) = \sigma_{N-1}^2. \quad \blacksquare$$

2.2. Tamaños de muestra y errores de estimación

2.2.1. Tamaños de muestra para la estimación de una media y una proporción

Los intervalos de confianza del capítulo anterior se basaron en el clásico teorema del límite central, el cual asume una muestra aleatoria de la variable en estudio. Desafortunadamente, en un MASs, que es a la larga el esquema de muestreo más utilizado, esta suposición no es correcta debido a la no independencia entre las componentes de las variables dadas en la proposición 2.1. Para subsanar este problema tenemos aquí dos caminos que dependerán de la naturaleza del tamaño de la muestra. Cuando esta es fija y el tamaño de la población $N \to \infty$, el esquema MASs converge en un MASc. Por otro lado, si $n \to \infty$, deberíamos también consentir que $N \to \infty$. Denotemos por μ_N y σ^2_{N-1} a la media y varianza de las correspondientes superpoblaciones. Hajek (1960) propuso el siguiente teorema del límite central: Si $\frac{n}{N} \to \tau \in]0,1[$ y $\text{máx}_{1\leq i\leq N} \frac{Y_i-\mu_N}{\sum_{i=1}^N (Y_i-\mu_N)^2} \to 0$ conforme $n \to \infty$ y $N \to \infty$ o $N\,\text{máx}_{1\leq i\leq N} \frac{Y_i-\mu_N}{\sum_{i=1}^N (Y_i-\mu_N)^2}$ es acotado en el límite cuando $N \to \infty$, entonces

$$Z = \frac{\bar{Y} - \mu_N}{\sqrt{1-\frac{n}{N}}\frac{\sigma_{N-1}}{\sqrt{n}}} \xrightarrow{\mathcal{D}} N(0,1),$$

conforme n y $N-n$ tiendan a infinito.

Este teorema del límite central nos permite entonces construir, utilizando como variable pivote la v.a. Z, un intervalo de confianza aproximado al $100(1-\alpha)\,\%$ para la media poblacional μ. Este, al suprimirse el subíndice $N-1$ en la varianza, toma para un tamaño de muestra y población suficientemente grandes la forma

$$IC = [\bar{Y} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\ ,\ \bar{Y} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}] = [\bar{Y} - z_{1-\frac{\alpha}{2}}SE(\bar{Y})\ ,\ \bar{Y} + z_{1-\frac{\alpha}{2}}SE(\bar{Y})\ ],$$

donde a $SE(\bar{Y})$, que es la raíz de la varianza asintótica de $\bar{Y}$, se le denomina el error estándar de estimación de $\bar{Y}$. Observe que este IC para μ difiere del clásico para poblaciones infinitas solo por el factor $\sqrt{1-\frac{n}{N}}$. Note además que si $N \to \infty$, este factor tiende a 1 y, por tanto, uno obtiene el clásico IC para μ.

De manera similar, es posible realizar un estudio inferencial para poblaciones finitas con una proporción poblacional p, ya que este es un caso particular de media cuando la variable Y es dicotómica. En este caso, la variable pivote Z normal toma la forma

$$Z = \frac{\bar{p} - p}{\sqrt{1-\frac{n}{N}}\sqrt{\frac{Np(1-p)}{n(N-1)}}},$$

con $\bar{p}$ igual a la proporción muestral, desde que $\sigma^2_{N-1} = \frac{Np(1-p)}{N-1}$. Así, si tomamos simétrica-

mente valores $-z_{1-\frac{\alpha}{2}}$ y $z_{1-\frac{\alpha}{2}}$ en la tabla normal estándar, podemos escribir:

$$P(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{p}-p}{\sqrt{1-\frac{n}{N}}\sqrt{\frac{Np(1-p)}{n(N-1)}}} \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha.$$

A fin de despejar p en esta expresión, podemos considerar la probabilidad equivalente

$$P(|\frac{\bar{p}-p}{\sqrt{1-\frac{n}{N}}\sqrt{\frac{Np(1-p)}{n(N-1)}}}|^2 \leq z^2_{1-\frac{\alpha}{2}}) = 1-\alpha$$

o

$$P(p^2(1+a) - p(2\bar{p}+a) + \bar{p}^2 \leq 0) = 1-\alpha,$$

donde $a = z^2_{1-\frac{\alpha}{2}}\frac{N-n}{n(N-1)}$. Esta probabilidad puede escribirse como

$$P((p-p_1)(p-p_2) \leq 0) = 1-\alpha,$$

siendo p_1 y p_2 las raíces de la ecuación asociada a la inecuación cuadrática anterior. Consecuentemente, $[p_1, p_2]$ constituye un IC tipo Wilson al $100(1-\alpha)\,\%$ para p. Si ahora en el IC anterior despreciamos el término $\frac{z^2_{1-\frac{\alpha}{2}}}{n}$, por ser este pequeño cuando n es grande, obtendremos el $IC = [p_1, p_2]$ al $100(1-\alpha)\,\%$ para p tipo Wald siguiente:

$$IC = [\bar{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\sqrt{\frac{N-n}{N-1}}\ ,\ \bar{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\sqrt{\frac{N-n}{N-1}}].$$

Si bien en el curso utilizaremos por simplicidad este último IC, hay que tener la precaución de que si la verdadera proporción es extrema (cercana a 0 o 1), este IC tipo Wald no presenta en general una adecuada cobertura. En tales situaciones, una opción más recomendable sería usar el IC tipo Wilson. Tal problema de cobertura puede ilustrarse a través del siguiente estudio de simulación, donde hemos graficado la proporción de cuántos de los 1000 IC, generados a través de 1000 MASs de tamaño 30 de una población de tamaño 400 contienen al verdadero parámetro p.

```
IC<-function(x,alpha,n,N,tipo){ # tipo = 1: Wald, tipo 2 = Wilson
        pbar = x/n
        z= qnorm(1-alpha/2)
        a = (z^2)*(N-n)/(n*(N-1))
        aux = a
        if(tipo==1) aux = 0
        e = 4*a*pbar + aux^2 - 4*a*pbar^2
        L1 = (2*pbar + aux - sqrt(e))/(2*(1+aux))
        L2 = (2*pbar + aux + sqrt(e))/(2*(1+aux))
        c(L1,L2)}
```

```
# Estudio de simulación:
cover <- function(n,N,p,alpha,tipo) {
        nsim = 1000
        count = 0
        for (i in 1:nsim) {
                x = rhyper(1,N*p,N*(1-p),n)
                if(tipo==1){ci = IC(x,alpha,n,N,1)}
                else {ci = IC(x,alpha,n,N,2)}
                if(p >= ci[1] & p <= ci[2]) {count = count + 1}
        }
        cover = count/nsim
        cover}
p = seq(0.005,0.995,by=0.01)
np = length(p)
cc1 = 0
cc2 = 0
N = 400
n = 30
for(j in 1:np){cc1[j] = cover(n,N,p[j],0.05,1)}
for(j in 1:np){cc2[j] = cover(n,N,p[j],0.05,2)}
```

Establecidas las fórmulas de los IC aproximados al $100(1-\alpha)\,\%$ para cualquier media y proporción poblacional, nos interesará ahora hallar el tamaño de muestra n que uno debería considerar para poder garantizar a un nivel de confianza del $100(1-\alpha)\,\%$ un error máximo de estimación e, donde por error de estimación entenderemos la diferencia en valor absoluto $|\hat{\theta}_n - \theta|$ entre el parámetro y su estimador. Esto se obtiene directamente de los IC obtenidos. En efecto, si queremos estimar μ, su IC correspondiente al $100(1-\alpha)\,\%$ puede reescribirse como

$$P(|\bar{Y} - \mu| \leq z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}) = 1 - \alpha.$$

Luego, según lo convenido, se debe tener que

$$e = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}},$$

de donde despejando obtenemos la siguiente fórmula para el tamaño de muestra:

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \sigma^2 N}{z^2_{1-\frac{\alpha}{2}} \sigma^2 + e^2 N}.$$

Note que si $N \to \infty$:

$$n = \frac{(z_{1-\frac{\alpha}{2}} \sigma)^2}{e^2}.$$

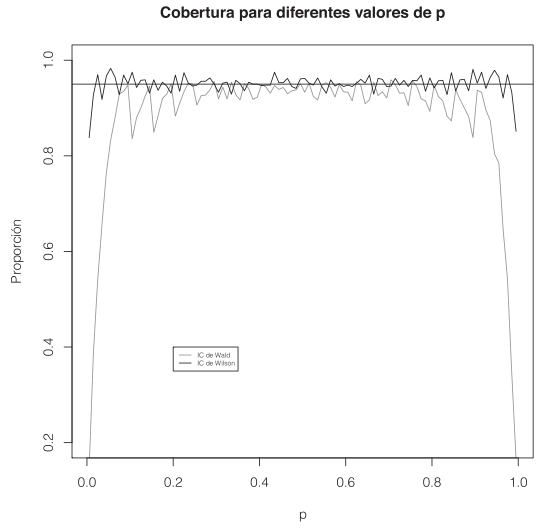


Figura 2.1: Simulación de la cobertura de los IC de Wald y Wilson al 95 % sobre una proporción

De manera similar, podemos deducir la siguiente fórmula del tamaño de muestra n para la estimación de p con un error máximo de estimación de e y un nivel de confianza del $100(1-\alpha)\,\%$:

$$n = \frac{(z^2_{1-\frac{\alpha}{2}}\bar{p}(1-\bar{p}))N}{z^2_{1-\frac{\alpha}{2}}\bar{p}(1-\bar{p}) + e^2(N-1)},$$

y si $N \to \infty$:

$$n = \frac{z^2_{1-\frac{\alpha}{2}}\bar{p}(1-\bar{p})}{e^2}.$$

Cabe agregar que la consideración de tamaños de muestra sobre la base de los errores máximos de estimación prefijados, también llamados errores absolutos e, no es universal. En la literatura es también común encontrar la consideración del coeficiente de variación o de los errores relativos. Recordemos que el coeficiente de variación poblacional (CV) de una variable estadística y se define como el cociente entre la desviación estándar y la media de esta variable, siendo este cociente usualmente expresado en porcentajes. La adimensionalidad de este indicador facilita claramente la determinación de valores objetivos sin que interese la escala en que uno mida la variable. Una regla práctica (que se debe de tomar con precaución) nos dice que un estimador no es confiable si su CV estimado supera 30 %; contrariamente, estimadores con un CV del 10 % o menos se suelen catalogar como confiables. Otra cantidad citada en el cálculo del tamaño de muestra es el error relativo, el cual se define como

$$e_r = z_{1-\frac{\alpha}{2}} CV(\hat{\theta}),$$

siendo $\hat{\theta}$ el estimador de interés para θ. Para su interpretación, basta notar que si $\hat{\theta}$ es un estimador insesgado y la muestra es suficientemente grande, tendremos que aproximadamente, con una confianza del $100(1-\alpha)\,\%$:

$$P(|\hat{\theta}-\theta| \leq z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{\theta})}) = 1-\alpha$$

o

$$P(|\frac{\hat{\theta}-\theta}{\theta}| \leq z_{1-\frac{\alpha}{2}}\frac{\sqrt{V(\hat{\theta})}}{E(\hat{\theta})}) = P(|\frac{\hat{\theta}-\theta}{\theta}| \leq z_{1-\frac{\alpha}{2}}CV(\hat{\theta})) = P(|\frac{\hat{\theta}-\theta}{\theta}| \leq e_r) = 1-\alpha.$$

Así, todas las fórmulas dadas en esta sección sobre n se satisfacen si, en lugar de especificarse e, uno especifica un error relativo e_r o un coeficiente de variación CV_0 para el estimador de interés a través de la siguiente relación:

$$e = \theta e_r = \theta z_{1-\frac{\alpha}{2}} CV_0.$$

2.2.2. Estimaciones previas

Un aspecto problemático en las fórmulas desarrolladas lo constituyen tanto σ como $\bar{p}$, ya que el primero es en general un parámetro poblacional no conocido y el otro no puede calcularse sin la muestra. En la práctica se tienen las siguientes alternativas para solucionar este problema:

- Estimar estas cantidades mediante un muestreo piloto (es decir, con una réplica previa, pero en escala menor del muestreo final).
- Estimar estas por cantidades similares de otros estudios semejantes.
- Estimar σ por $\hat{\sigma} = \frac{Rango}{6}$, donde *Rango* denota el ancho del intervalo que estimamos contiene a todos los posibles valores de la variable Y. Esto se justifica por la desigualdad de Chebyshev, la cual, recordemos, nos dice que la probabilidad de que Y se encuentre en el intervalo $[\mu-3\sigma, \mu+3\sigma]$, siendo μ la media de Y, es muy cercana a 1 (concretamente de al menos 0.89).
- Tomar $\bar{p} = \frac{1}{2}$. Esta es una regla conservadora que simplemente asigna el valor de $\bar{p}$, que maximiza el tamaño de muestra. Así, uno podrá siempre garantizar, al margen del verdadero $\bar{p}$, un error de estimación de a lo más e.

Ejemplo 2.2. *La facultad de Ingeniería de una universidad cuenta con 1200 alumnos y está interesada en realizar una encuesta con el fin de determinar, entre otros, el número de alumnos que tienen una PC en casa. El coordinador de la facultad desea estimar este*

total con un error máximo no mayor a los 30 alumnos y una confianza del 95 % ¿A cuántos alumnos de la facultad se les debería aplicar la encuesta?
Solución: Se desea estimar τ = número de alumnos de la facultad que poseen una PC en casa con un margen de error no mayor a los 30 alumnos y un nivel de confianza del 95 %. Dado que la población de alumnos en la facultad es finita ($N = 1200$) y $\tau = Np$, donde p denota la proporción de alumnos de la facultad que poseen una PC en casa, el problema equivale a estimar p con un margen de error no mayor a $e = \frac{30}{1200} = 0.025$ y un nivel de confianza del 95 %. Dado que $\bar{p}$ se desconoce, tomaremos la regla conservadora $\bar{p} = \frac{1}{2}$. Con ello, el tamaño de muestra requerido será de

$$n = \frac{z^2_{0.975} \times 0.5^2 \times 1200}{z^2_{0.975} \times 0.5^2 + 0.025^2 \times 1199} = 674.0536 \simeq 675 \quad \textit{alumnos.}$$

Cabe observar que de no haberse tomado en cuenta el tamaño de la población ($N \to \infty$), uno hubiese obtenido, bajo el mismo error de estimación de 0.025, un tamaño de muestra de $n = 1537$ alumnos, lo cual ciertamente no tiene sentido. □

Observaciones

- Dado que los tamaños de muestra se han basado en el estudio de un solo parámetro, es lógico preguntarse qué pasaría si en una investigación existen varios parámetros o variables de interés. En tal caso se sugiere ubicar, según los objetivos del estudio, cuáles son los parámetros de relevancia. Hecho esto, uno puede obtener tantos tamaños de muestra como parámetros de interés tenga y tomar el mayor valor de estos. Tal estrategia garantiza que en todos los casos relevantes uno obtenga a lo más los errores de estimación pre establecidos.

- Los tamaños de muestra calculados deben siempre aproximarse por exceso a un número entero; de lo contrario, no satisfaceríamos el requerimiento del máximo error preestablecido. Por otro lado, en la práctica es importante inflar estos tamaños por no respuesta. La información de tasas de no respuesta en estudios previos, pilotos o similares es en muchas situaciones fácil de obtener.

- Hemos priorizado en el curso el muestreo bajo el contexto que nos interesa estimar ciertos parámetros poblacionales. Sin embargo, en algunas aplicaciones el estudio es comparativo o correlacional y más que estimar puntualmente parámetros con una precisión determinada nos podría interesar, por ejemplo, poder detectar ciertas diferencias entre las medias o proporciones de las poblaciones a comparar o estimar el efecto de ciertas variables en un análisis de regresión. Estos análisis estadísticos bajo un MASs, o más genéricamente bajo una muestra compleja, se introducirán en el capítulo 5.

2.3. Aspectos computacionales y el paquete survey

Existen en la literatura diferentes softwares estadísticos que pueden utilizarse para analizar muestras complejas. Información sobre estos puede encontrarse, por ejemplo, en

`http://www.hcp.med.harvard.edu/statistics/survey-soft/.`

Nosotros usaremos, aparte del siempre útil Excel y de ciertas rutinas de R, los paquetes survey y sampling de R. Del segundo nos ocuparemos en los capítulos posteriores. En cuanto al primero, este tiene esencialmente dos propósitos principales:

- Enlazar la data al diseño de metadata (pesos, probabilidades de selección, unidades primarias, identificadores de estratos, etc.) con el fin de poder realizar los ajustes que sean necesarios al diseño de manera confiable y automática. Esto se hace con las funciones `svydesign` y `svrepdesign` que crean objetos que contienen no solo la base de datos, sino también la información del diseño. Así, por ejemplo, uno podría extraer un subconjunto de la data y preservar su diseño aplicado a este subconjunto.

- Proveer de estimaciones válidas, con sus errores estándar de estimación estimados, para diferentes estadísticos y procedimientos, de tal manera que se respete el diseño de muestreo probabilístico empleado.

El primer paso para realizar un análisis con el paquete survey consiste en crear un objeto diseño apropiado que contenga la data y la metada necesarias. Esto se hace con la función `svydesign` o `svrepdesign` en caso de que se den pesos de replicación. Las funciones de análisis usualmente toman como argumento el objeto diseño y una fórmula modelo que especifica las variables que se usarán. Los nombres de las funciones de análisis para los objetos creados con `svydesign` y `svrepdesign` comienzan con `svy` y `svr,` respectivamente.

Seguidamente brindaremos una introducción al uso del paquete survey y de paso presentaremos algunas bases reales de datos censales que utilizaremos a lo largo del curso.

2.3.1. La base de datos api

Nuestro primer ejemplo del uso del paquete survey será con el análisis de un MAS para la población contenida en la base de datos `api`. Una descripción de esta base junto y la información de las 37 variables en ella consideradas puede encontrarse en

`http://cran.fhcrc.org/web/packages/survey/survey.pdf`

Cabe comentar, como resumen, que el estado de California exige anualmente una evaluación de sus escuelas públicas. En tal sentido, el departamento de educación de este estado registra anualmente el índice api (de academic performance index), que mide cuán bien va una escuela

en términos de rendimiento. El archivo api contiene este índice e información demográfica de todas las 6194 escuelas públicas de California con al menos 100 alumnos por escuela.

Para acceder a la base de datos y al uso del paquete survey (que debe instalarse con antelación) escribamos

```
library(survey)
data(api)
head(apipop,4)

##              cds stype            name                     sname snum
## 1 01611190130229     H    Alameda High              Alameda High    1
## 2 01611190132878     H    Encinal High              Encinal High    2
## 3 01611196000004     M  Chipman Middle             Chipman Middle    3
## 4 01611196090005     E Lum (Donald D.) Lum (Donald D.) Elementary    4
##                   dname dnum   cname cnum flag pcttest api00 api99 target
## 1 Alameda City Unified    6 Alameda    1   NA      96   731   693      5
## 2 Alameda City Unified    6 Alameda    1   NA      99   622   589     11
## 3 Alameda City Unified    6 Alameda    1   NA      99   622   572     11
## 4 Alameda City Unified    6 Alameda    1   NA      99   774   732      3
##   growth sch.wide comp.imp both awards meals ell yr.rnd mobility acs.k3
## 1     38      Yes      Yes  Yes    Yes    14  16   <NA>        9     NA
## 2     33      Yes       No   No     No    20  18   <NA>       13     NA
## 3     50      Yes      Yes  Yes    Yes    55  25   <NA>       20     NA
## 4     42      Yes      Yes  Yes    Yes    35  26   <NA>       21     20
##   acs.46 acs.core pct.resp not.hsg hsg some.col col.grad grad.sch avg.ed
## 1     NA       25       91       6  16       22       38       18   3.45
## 2     NA       27       84      11  20       29       31        9   3.06
## 3     26       27       86      11  31       30       20        8   2.82
## 4     30       NA       96       3  22       29       31       15   3.32
##   full emer enroll api.stu
## 1   85   16   1278    1090
## 2   90   10   1113     840
## 3   80   12    546     472
## 4   96    4    330     272
```

Aquí mostramos los cuatro primeros registros de la base de datos `api` (que está en `apipop`) . La primera línea de comandos solo fija el directorio de trabajo. Consideremos ahora un MASs de escuelas públicas de tamaño 100, donde hemos fijado la semilla aleatoria anteriormente comentada para que usted pueda replicar los mismos resultados aquí obtenidos.

```
set.seed(12345)
N = dim(apipop)[1]
n = 100
index1 = sample(N,n)
sample1 = apipop[index1,]
```

Por razones que comentaremos luego, será también interesante agregar a esta base dos nuevas variables: fpc y pp. La primera es el tamaño de la población (6194); y la otra, la probabilidad de selección de cada elemento en la población $pp = \frac{n}{N}$. Ello se hace con

```
aux = data.frame(fpc = rep(N,100), pp = rep(n/N,100))
sample1 = cbind(sample1,aux)
```

Definamos ahora un objeto diseño apropiado que contenga la data y metada necesarias. Esto se hace con

```
diseMASs = svydesign(ids = ~1,fpc = ~fpc,data = sample1)
```

El argumento `ids` es para indicar las variables de conglomeración, las cuales en nuestro caso no existen y es por ello que colocamos `ids`= 1. El argumento `fpc` (de factor de corrección para poblaciones finitas) indica el tamaño de la población, con lo cual implícitamente asumimos que se deben aplicar las formulaciones de corrección para poblaciones finitas y que se está realizando un muestreo sin reemplazamiento. La notación $\sim$ indica que la variable `fpc` está ya definida en la muestra `sample1`. Si el argumento `fpc` se omite, entonces deben indicarse las probabilidades de selección o los pesos de muestreo, en cuyo caso se estaría asumiendo implícitamente un muestreo con reemplazamiento. Tanto `ids` como `fpc`, aparte de los valores por defecto, conforman la metadata del diseño.

Otro diseño que podría aplicarse en este mismo ejemplo es un MASc, para lo cual deberíamos tomar formalmente la muestra aleatoria con reemplazamiento mediante

```
set.seed(12345)
sample2 = apipop[sample(N,100, replace=TRUE),]
sample2 = cbind(sample2,aux)
```

El objeto diseño correspondiente sería

```
diseMASc = svydesign(ids = ~1,probs = ~pp,data = sample2)
```

De pedirse información, obtendríamos

```
diseMASc

## Independent Sampling design (with replacement)
## svydesign(ids = ~1, probs = ~pp, data = sample2)
```

Supongamos ahora que estamos interesados en estimar ciertos parámetros poblacionales, como, por ejemplo, el número total de alumnos matriculados, la proporción por tipo de escuelas y las medias y diferencia de medias del api entre 1999 y 2000. Bajo el diseño MASs, esto se puede hacer mediante:

```
svytotal(~enroll,diseMASs)

##        total SE
## enroll    NA NA

svymean(~stype, diseMASs)

##        mean   SE
## stypeE 0.68 0.05
## stypeH 0.20 0.04
## stypeM 0.12 0.03

means1 = svymean(~api00+api99,diseMASs)
means1

##       mean   SE
## api00  652 12.6
## api99  628 12.9

svycontrast(means1,c(api00=1,api99=-1))

##          contrast   SE
## contrast     24.5 2.96
```

El hecho de que en el primer resultado se obtenga NA se debe a que en la muestra existe en la muestra algún o algunos casos perdidos. Esto puede corregirse eliminando tales mediante

```
svytotal(~enroll,diseMASs,na.rm=T)

##          total     SE
## enroll 4115727 291390
```

Con un MASc, lo anterior se convierte en

```
svytotal(~enroll,diseMASc,na.rm=T)

##          total     SE
## enroll 3979335 303578
```

Tenemos también

```
svymean(~stype, diseMASc)

##        mean   SE
## stypeE 0.70 0.05
## stypeH 0.11 0.03
## stypeM 0.19 0.04

(means1 = svymean(~api00+api99,diseMASc))

##       mean   SE
## api00  678 11.6
## api99  648 12.1

svycontrast(means1,c(api00=1,api99=-1))

##          contrast   SE
## contrast     30.4 2.84
```

2.3.2. La evaluación censal de estudiantes 2019

La unidad de medición de la calidad de los aprendizajes (UMC) del Ministerio de Educación, publicó en 2020 los resultados de la última evaluación censal de estudiantes (ECE) 2019. La página web correspondiente contiene información variada, entre la que destacan las bases de datos en formato SPSS no solo de la ECE 2019 sino también la de años anteriores. Nosotros consideraremos inicialmente a la población objetivo de los rendimientos en el segundo grado de secundaria de la Dirección Regional de Amazonas (en adelante DRE Amazonas). Más adelante trabajaremos con una población mayor. Vale reiterar que estos datos son censales, aunque en el caso del segundo grado se incluyen solo a aquellas escuelas con más de 5 alumnos. No estamos tampoco incluyendo los factores de ajuste o ponderación por casos pérdidos, que se incluyen para replicar los resultados dados por la UMC. Las variables de interés para esta base de datos serán los puntajes de evaluación en las áreas de

Lectura, Matemáticas y Ciencia y Tecnología (todas en una escala Rasch normalizada a 500 puntos). Para el Ministerio, los niveles de logro son de particular interés. Estos se obtienen al categorizar los puntajes anteriores en cuatro niveles: previo al inicio, en inicio, en proceso y satisfactorio.

Luego de instalar el paquete `foreign`, podremos operacionalizar las bases de datos nacional mediante

```
library(foreign)
ece19 = read.spss(file.choose(), to.data.frame=TRUE)
#file.choose() busca en su hardware el archivo SPSS ECE_2S_2019_WEB.sav
setwd("~/Documents/TextoMuestreo2020")
# setwd fija el directorio de trabajo (DT)
save(ece19,file='ece19.RData')
#El archivo ece19.RData se grabará mediante save en su DT
```

Se muestran abajo, los primeros tres registros de la base de la DRE Amazonas

```
setwd("~/Documents/TextoMuestreo2020")
load("ece19.RData")
ece19Am = ece19[ece19$Departamento==levels(ece19$Departamento)[1],]
#save(ece19Am,file='ece19Am.RData')
head(ece19Am,3)

##       ID_IE ID_Seccion cor_est cod_DRE            nom_dre cod_UGEL
## 44817 21273         01      01    0100 Amazonas             010002
## 44818 21273         01      02    0100 Amazonas             010002
## 44819 21273         01      03    0100 Amazonas             010002
##                             nom_ugel codgeo  Departamento
## 44817 Bagua                          010201 AMAZONAS
## 44818 Bagua                          010201 AMAZONAS
## 44819 Bagua                          010201 AMAZONAS
##                       Provincia                              Distrito
## 44817 BAGUA                    BAGUA
## 44818 BAGUA                    BAGUA
## 44819 BAGUA                    BAGUA
##       gestion2   area   sexo M500_L    grupo_L M500_M    grupo_M M500_CT
## 44817  Estatal Urbana Hombre    639 En proceso    620 En proceso     542
## 44818  Estatal Urbana Hombre    634 En proceso    647 En proceso     602
## 44819  Estatal Urbana Hombre    616 En proceso    563  En inicio     620
##         grupo_CT aj_lectura aj_matematica aj_ct    ISE
```

```
## 44817 En proceso       1.03          1.03  1.07 -0.849
## 44818 En proceso       1.03          1.03  1.07  0.826
## 44819 En proceso       1.03          1.03  1.07  0.928
```

Note que, a diferencia de la base de datos api, las unidades en esta base son alumnos y no colegios.

Supongamos ahora que nuestro interés sea estimar el rendimiento medio de los alumnos tanto en Lectura (L), Matemáticas (M) y Ciencia y Tecnología (CT), con un margen de error no mayor a 5 puntos y un nivel de confianza del 95 %. Para encontrar el tamaño de muestra requeriremos de estimaciones de la varianza de estos puntajes, las cuales las podríamos obtener de la ECE 2018 o a través de un estudio piloto. Si optamos por un piloto de 30 alumnos, la selección correspondiente, así como la estimación de las varianzas requeridas, se hará como sigue.

```
set.seed(12345)
N = dim(ece19Am)[1]
index1 = sample(N,30)
mp19Am = ece19Am[index1,]
dismp = svydesign(id=~1,fpc=rep(N,30),data=mp19Am)
sigmae2_L = coef(svyvar(~M500_L,dismp,na.rm=T))
sigmae2_M = coef(svyvar(~M500_M,dismp,na.rm=T))
sigmae2_CT = coef(svyvar(~M500_CT,dismp,na.rm=T))
```

Dado que tenemos tres variables, optaremos, como comentamos, por seleccionar el mayor tamaño de muestra bajo estas utilizando un redondeo por exceso.

```
d = 25*N/(qnorm(0.975)^2)
n1 = N*sigmae2_L/(d + sigmae2_L)
n2 = N*sigmae2_M/(d + sigmae2_M)
n3 = N*sigmae2_CT/(d + sigmae2_CT)
(n = ceiling(max(n1,n2,n3)))

## [1] 1662
```

La toma de muestra, definición del diseño y estimaciones de los rendimientos y proporciones de logro se muestran a continuación:

```
set.seed(12345)
index = sample(N,n)
m19Am = ece19Am[index,]
```

```
disem = svydesign(id=~1,fpc=rep(N,n),data=m19Am)
svymean(~M500_L,disem,na.rm=T)

##        mean   SE
## M500_L  536 1.62

svymean(~M500_M,disem,na.rm=T)

##        mean   SE
## M500_M  533 2.29

svymean(~M500_CT,disem,na.rm=T)

##         mean   SE
## M500_CT  469 2.59

meanp_L = svymean(~grupo_L,disem,na.rm=T)
meanp_M = svymean(~grupo_M,disem,na.rm=T)
meanp_CT = svymean(~grupo_CT,disem,na.rm=T)

pr = rbind(meanp_L,meanp_M,meanp_CT)
colnames(pr) = c("Previo al inicio","Inicio","En proceso","Satisfactorio")
pr

##          Previo al inicio Inicio En proceso Satisfactorio
## meanp_L             0.331  0.403      0.187        0.0790
## meanp_M             0.448  0.294      0.145        0.1133
## meanp_CT            0.209  0.410      0.313        0.0675
```

2.3.3. El censo nacional de población penitenciaria 2016

El censo nacional de población penitenciaria 2016, realizado por primera vez en el país por el Instituto Nacional de Estadística e Informática (INEI), generó información estadística cuantitativa y cualitativa actualizada sobre la problemática penitenciaria en el Perú. La base de datos de este censo es de libre disponibilidad y se puede encontrar en la siguiente página web del INEI:

`http://iinei.inei.gob.pe/microdatos/`.

La versión de esta base de datos, que utilizaremos a lo largo del curso, se encuentra en el archivo BasR.sav. Está en formato SPSS y contiene todos los 76 180 registros de personas

privadas de libertad en el país consignadas en el censo y la gran mayoría de preguntas de la encuesta, la cual también se encuentra disponible en la página web del INEI. Para utilizar la base de datos en R, debemos instalar el paquete foreign y luego invocar los comandos

```
library(foreign)
#cp16b <- read.spss(file.choose(), use.value.labels=TRUE)
cp16b <- read.spss("BasR.sav", use.value.labels=TRUE)
cp16 = as.data.frame(cp16b)
cp16_labels <- attr(cp16b, "variable.labels")
cp16_cat <- attr(cp16b, "label.table")
save(cp16,file='cp16.RData')
```

La base de datos a utilizar es cp16; mientras que los archivos `cp16_labels` y `cp16_cat` contienen información de, respectivamente, las etiquetas y categorías de las variables seleccionadas. Como se aprecia, la base de datos cp16 ha sido también grabada para uso futuro en el formato de R. Esta base tiene, como seguidamente se aprecia, 189 variables, de las cuales mostramos las primeras 18.

```
head(cp16[,1:18])

##   ID      PDEP      PPROV       PDIS        PCP
## 1  3 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
## 2 19 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
## 3 24 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
## 4 26 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
## 5 39 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
## 6 40 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
##                          OFICINA_R EST_PENIT PABELLON GENERO     E_CIVIL
## 1 Oficina Regional Norte Chiclayo Cajamarca        4  Mujer   Casado(a)
## 2 Oficina Regional Norte Chiclayo Cajamarca       NA  Mujer    Viudo(a)
## 3 Oficina Regional Norte Chiclayo Cajamarca       NA Hombre   Casado(a)
## 4 Oficina Regional Norte Chiclayo Cajamarca       NA Hombre    Viudo(a)
## 5 Oficina Regional Norte Chiclayo Cajamarca        3 Hombre   Casado(a)
## 6 Oficina Regional Norte Chiclayo Cajamarca        7 Hombre Conviviente
##     RELIGION EDAD    NACIONALIDAD        PAIS_NAC     DEP_NAC  DEP_URES
## 1   Católica   39 PERUANO         PERU                   LIMA      LIMA
## 2     Mormón   49 PERUANO         PERU                   LIMA      LIMA
## 3    Ninguna   25 PERUANO         ESTADOS UNIDOS           NA        NA
## 4       Otra   26 PERUANO         PERU                  CUSCO      LIMA
## 5 Evangélica   49 PERUANO         PERU               CAJAMARCA CAJAMARCA
```

```
## 6    Ninguna   40 PERUANO         PERU            LA LIBERTAD CAJAMARCA
##                CP_URES                                DEL_GENERICO_CD
## 1 CIUDAD DE DIOS                     DELITOS CONTRA EL PATRIMONIO
## 2 BARRIO OBRERO INDUST               DELITOS CONTRA EL PATRIMONIO
## 3                                    DELITOS CONTRA EL PATRIMONIO
## 4 VILLA EL SALVADOR                  DELITOS CONTRA EL PATRIMONIO
## 5 LA COLPA              DELITOS CONTRA LA ADMINISTRACION PUBLICA
## 6 CAJAMARCA                          DELITOS CONTRA EL PATRIMONIO
```

La distribución de frecuencias del número de internos, condición de género (CG) y capacidad de los establecimientos penitenciarios en cada oficina regional y departamento se muestran en el cuadro 2.4.

Como una primera aproximación al análisis de la base de datos consideraremos un MASs, cuyo objetivo será estimar cualquier proporción poblacional con un margen de error no mayor a 0.03 y una confianza del 95 %. Para ello, el tamaño de muestra requerido estará dado por

$$n = \frac{1.96^2 \times 0.5 \times (1 - 0.5) \times 76\,180}{1.96^2 \times 0.5 \times (1 - 0.5) + 0.03^2 \times 76\,179} = 1052.383$$

que redondeando nos da un valor de 1053 internos. Si bien usaremos este número, cabe comentar que ello es si asumimos que todos responderán a la encuesta. En encuestas similares para la región se han encontrado tasas de no respuesta de entre el 21 y 22 %. Una práctica que comentamos es la de inflar este número ante la posibilidad de no respuesta. Ello nos sugeriría encuestar a 1285 internos. Para efectos de este ejercicio tomaremos solo 1053, ya que en nuestro caso es posible acceder a toda la información. Tomada la muestra, estimemos la edad promedio de los internos, la proporción de internos sentenciados y la proporción de estos que tienen un abogado. Los códigos siguientes nos permitarán hacer todo ello.

```
set.seed(12345)
load('cp16.RData')
N = dim(cp16)[1]
index = sample(N,1053)
sample = cp16[index,]
diseMASs = svydesign(id=~1,fpc=rep(N,1053),data = sample)
svymean(~EDAD, diseMASs,na.rm=T)

##      mean   SE
## EDAD 35.8 0.35

svymean(~SITUACION_JURIDICA,diseMASs,na.rm=T)
```

OFICINA REGIONAL	DEPARTAMENTO	E.PENITENCIARIO	NUMERO DE INTERNOS	CG	Capacidad
Norte Chiclayo	CAJAMARCA	Cajamarca	1389	Mix	888
		Chota	131	H	65
		Jaen	377	Mix	50
		San Ignacio	79	H	150
	LA LIBERTAD	Pacasmayo	11	M	72
		Trujillo	4471	H	1518
		Mujeres de Trujillo	283	M	160
	LAMBAYEQUE	Chiclayo	3163	Mix	1143
	PIURA	Piura	3098	H	1370
		Sullana	94	M	50
	TUMBES	Tumbes	860	Mix	384
Lima	ANCASH	Huaraz	1014	Mix	350
		Chimbote	2321	Mix	920
	CALLAO	Callao	3201	H	572
		Base Naval Callao	7	H	8
	ICA	Chincha	1331	H	1152
		Ica	3943	Mix	1464
	LIMA	Cañete	1982	H	768
		Huaral	3164	H	823
		Huacho	1738	Mix	644
		Ancon	2289	H	1620
		Modelo Ancon II	1462	Mix	2200
		Anexo Mujeres Chorrillos	309	M	288
		Mujeres de Chorrillos	810	M	450
		Virgen de Fatima	339	M	548
		Virgen de la Merced	13	H	42
		Lurigancho	9602	H	3204
		Miguel Castro Castro	4359	H	1142
		Barbadillo	1	H	1
Sur Arequipa	AREQUIPA	Arequipa	1971	H	667
		Mujeres de Arequipa	151	M	67
		Camana	262	H	78
	TACNA	Tacna	830	H	222
		Mujeres de Tacna	110	M	40
		Challapalca	162	H	214
Centro Huancayo	AYACUCHO	Ayacucho	2438	Mix	644
		Huanta	101	H	42
	HUANCAVELICA	Huancavelica	200	H	60
	JUNIN	Chanchamayo	572	Mix	120
		Huancayo	1972	H	680
		Mujeres de Concepción	31	M	105
		Jauja	104	M	85
		Satipo	164	H	50
		Tarma	84	H	48
		Oroya	114	Mix	64
Oriente Huanuco (Pucallpa)	HUANUCO	Huanuco	2554	Mix	1074
	PASCO	Cerro Pasco	195	Mix	96
	UCAYALI	Pucallpa	2053	Mix	788
Sur Oriente Cusco	APURIMAC	Abancay	256	Mix	90
		Andahuaylas	354	Mix	248
	CUSCO	Cusco	2288	H	800
		Mujeres del cusco	137	M	62
		Quillabamba	347	Mix	80
	MADRE DE DIOS	Pto. Maldonado	712	H	590
Nor Oriente San Martín	AMAZONAS	Chachapoyas	629	Mix	288
		Bagua Grande	230	Mix	60
	LORETO	Yurimaguas	157	Mix	286
		Iquitos	1025	H	600
		Mujeres de Iquitos	64	M	78
	SAN MARTIN	Juanjui	686	Mix	654
		Moyobamba	588	Mix	544
		Sananguillo	548	H	636
		Tarapoto	463	H	180
Altiplano Puno	PUNO	Lampa	136	M	44
		Puno	582	H	778
		Juliaca	1069	Mix	420

Cuadro 2.4: Distribución de frecuencias del número de internos, condición de género (CG) y capacidad de los establecimientos penitenciarios en cada oficina regional y departamento del Perú

```
##                                  mean   SE
## SITUACION_JURIDICAProcesado   0.222 0.01
## SITUACION_JURIDICASentenciado 0.778 0.01

svymean(~ABOGADO,diseMASs,na.rm=T)

##           mean   SE
## ABOGADOSí 0.53 0.02
## ABOGADONo 0.47 0.02
```

Otro análisis de interés podría ser determinar si existe relación entre si el interno consumía drogas o no y el tipo de delito que ha cometido. Antes de analizar ello será conveniente recodificar la tipicidad del delito a los delitos más comunes, creando la variable DGEN. Como la prueba indica y se visualiza en el gráfico de barras agrupadas, no encontramos evidencia de una asociación entre el consumo de drogas y la tipificación del delito.

```
DGEN = cp16$DEL_GENERICO_CD
levels(DGEN)[c(1,2,3,4,5,7,8,9,10,11,14,16,17,18,19)] = "OTROS"
DGEN = DGEN[index]
DGEN = factor(DGEN,levels(DGEN)[c(2,3,4,5,1)])
chisq.test(DGEN,sample$DROGAS)

##
##  Pearson's Chi-squared test
##
## data:  DGEN and sample$DROGAS
## X-squared = 3, df = 4, p-value = 0.6

tab = table(sample$DROGAS,DGEN)
```

2.3.4. La población peruana con DNI 2018

Este último ejemplo considera a la población peruana que se encuentra en el Registro Nacional de Identificación y Estado Civil (RENIEC) al 31 de diciembre de 2018 y que, por tanto, cuenta con su documento nacional de identidad (DNI), el cual otorga derecho a sufragio a partir de los 18 años. La información pública del RENIEC incluye el lugar de residencia, edad, sexo y condición de extranjería de la persona. Esta base de datos puede obtenerse en formato Excel o SPSS desde la página web de esta institución. Una mirada a la base de datos

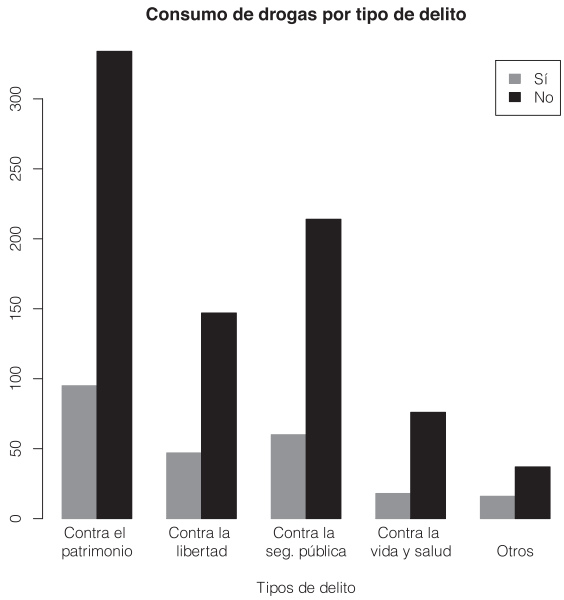


Figura 2.2: Frecuencias de consumo de drogas por tipo de delito

```
library(foreign)
reniec18 = read.spss("BD_Pob_Identificada_2018_Spss.sav")
reniec18 = data.frame(lapply(reniec18, trimws))
head(reniec18,3)

##   RESIDENCIA UBIGEO_RENIEC UBIGEO_INEI Continente_R CONTIO Pais_R PAIS_0
## 1   Nacional        010101      010101      América          Perú
## 2   Nacional        010101      010101      América          Perú
## 3   Nacional        010101      010101      América          Perú
##   DEPARTAMENTO   PROVINCIA    DISTRITO   SEXO EDAD CANTIDAD
## 1     Amazonas Chachapoyas Chachapoyas Hombre    0      336
## 2     Amazonas Chachapoyas Chachapoyas Hombre    1      366
## 3     Amazonas Chachapoyas Chachapoyas Hombre    2      361
```

revela que la última variable, CANTIDAD, contiene la frecuencia de casos que comparten las demás variables. Como ilustración, en 2018 se tenían 336 varones registrados en el distrito de Chachapoyas, provincia de Chachapoyas y departamento de Amazonas, que no cumplían aún el año de edad. Esta variable, por tanto, es una variable de ponderación para toda la base de datos, con lo cual ella contendrá a nivel nacional una cantidad de registros igual a

```
Cantidad = as.numeric(paste(reniec18$CANTIDAD))
(N = sum(Cantidad))

## [1] 34894246
```

En este ejemplo estaremos interesados en estimar cualquier proporción de interés con un margen de error no mayor a 0.02 y una confianza del 95 % mediante un MASs. Esto podría ser útil, por ejemplo, en una encuesta de opinión pública, solo que para acceder a la vivienda específica del entrevistado se debería conocer la dirección u otra información pertinente. Si tomamos la regla conservadora de $\bar{p} = 0.5$, entonces el tamaño de muestra requerido será de

```
(n = qnorm(0.975)^2*0.5^2*N/(qnorm(0.975)^2*0.5^2 + 0.02^2*(N-1)))

## [1] 2401
```

Para tomar esta muestra requeriremos expandir antes la base de datos de individuos por la variable CANTIDAD. Esto puede hacerse con el siguiente comando en R, en el cual generaremos la base de datos expandida reniec18x.RData:

```
reniec18x = reniec18[rep(1:nrow(reniec18),Cantidad),]
reniec18x = cbind(id=1:N,reniec18x)
save(reniec18x,file='reniec18x.RData')
```

Si tomamos el MASs planificado, obtendremos la siguiente base de datos muestral:

```
load('reniec18x.RData')
set.seed(12345)
indexp = sample(N,2401)
sampleDNI = reniec18x[indexp,]
sampleDNI[1:7,c(1,8:12)]

##                   id PAIS_0 DEPARTAMENTO         PROVINCIA  DISTRITO
## 267934.451  25155398               Lima            Huaral    Huaral
## 326175.16   30559446               Puno              Moho      Moho
## 280492.53   26553905             Loreto Datem del Marañón    Andoas
## 331922.32   30920649          San Martín  Mariscal Cáceres Pajarillo
## 240810.4190 15928559               Lima              Lima     Comas
## 125997.100   5805418          Cajamarca       San Ignacio  Namballe
## 216292.2041 11343959         La Libertad          Trujillo  Trujillo
```

```
##               SEXO
## 267934.451   Mujer
## 326175.16    Mujer
## 280492.53    Mujer
## 331922.32   Hombre
## 240810.4190 Hombre
## 125997.100  Hombre
## 216292.2041  Mujer
```

Si bien considerar aquí un MASs es teóricamente posible y ha sido en este y en los anteriores ejemplos bastante simple, este no es ciertamente un diseño recomendable para poblaciones tan grandes como las aquí consideradas. En nuestros ejemplos contamos en todos los casos con una base de datos poblacional, situación que raramente se presenta en la práctica. En la realidad, frecuentemente el marco muestral está desactualizado, pobremente definido o es inexistente y, por otro lado, la muestra aleatoria simple resulta estar tan geográficamente dispersa que los costos y la logística resultan inmanejables. En esta muestra, por ejemplo, apreciemos el lugar de residencia de las 6 primeras personas seleccionadas. Si la encuesta objetivo es de opinión y se puede incluso obtener la dirección de las personas a encuestar, demandaría un arduo y costoso trabajo tratar de ubicarlas por la lejanía entre ellas y el aparato logístico que se tendría que implementar para garantizar la supervisión y calidad del trabajo de campo. En los capítulos siguientes exploraremos diseños mucho más eficientes para los fines buscados.

Para terminar, obtengamos la estimación de la proporción de mujeres y de personas con derecho a votar (con 18 o más años de edad) en esta población.

```
diseDNI = svydesign(ids=~1,fpc=rep(N,nrow(sampleDNI)),data=sampleDNI)
Edad=as.numeric(paste(sampleDNI$EDAD))
diseDNI = update(diseDNI,Edad)
svymean(~Edad>=18,diseDNI)

##                   mean   SE
## Edad >= 18FALSE 0.309 0.01
## Edad >= 18TRUE  0.691 0.01
```

2.4. Ejercicios

1. Considere una población conformada por 6 personas, a las que se les ha medido el nivel de hemoglobina en gramos por decilitro, y en las que se ha encontrado las siguientes mediciones

$$13.9, 11.5, 16.7, 14.4, 14.6, 15.1.$$

Mediante un MASc y un MASs de tamaño $n = 3$,
a) Halle la probabilidad de que la media del nivel de hemoglobina de las 3 personas seleccionadas supere los 14 gramos por decilitro.
b) Suponga que para estimar el nivel promedio de hemoglobina en estas personas se propone la mediana de los valores observados en la muestra. ¿Sería este un estimador insesgado? ¿Tiene este una menor varianza que la media muestral?
c) Usando los números aleatorios 0.018, 0.310 y 0.549, tome las muestras requeridas y estime la media del nivel de hemoglobina de las 6 personas.

2. Una manera de estimar el tamaño N de una población consiste en usar métodos de captura-recaptura. Estos consisten en seleccionar primero al azar m elementos de la población para reponerlos en ella luego de marcarlos. Seguidamente se tienen dos estrategias. El método directo consiste en seleccionar al azar y sin reemplazamiento otra muestra de n elementos de la población para registrar luego el número de elementos marcados X que se encuentren en ella. El segundo método, llamado muestreo inverso, consiste en seleccionar al azar y con reemplazamiento (podría también analizar el caso sin reemplazamiento) sistemáticamente elementos de la población hasta ubicar r elementos marcados. Con ello se tienen los siguientes dos estimadores de N:

$$\hat{N}_1 = \frac{nm}{X} \quad \text{y} \quad \hat{N}_2 = \frac{mY}{r},$$

donde Y denota el número de intentos hasta obtener la cuota de r elementos marcados.
a) Usando una expansión de Taylor adecuada, muestre que aproximadamente se cumple que $E(\hat{N}_1) = N + \frac{2N(N-m)(N-n)}{nm(N-1)}$ y $V(\hat{N}_1) = \frac{N^2(N-m)(N-n)}{nm(N-1)}$.
b) Como se aprecia en a), $\hat{N}_1$ es no solo un estimador sesgado de N, sino que presenta una gran varianza si la muestra correspondiente contiene muy pocos elementos marcados. Muestre que, contrariamente, $\hat{N}_2$ es un estimador insesgado de N y que tiene una varianza igual a

$$V(\hat{N}_2) = \frac{N(N-m)}{r}.$$

Pruebe además que

$$\hat{V}(\hat{N}_2) = \frac{m^2 Y(Y-r)}{r^2(r+1)}$$

es un estimador insesgado de la varianza última. ¿Qué desventaja cree que pudiera tener este método con respecto al muestreo directo?

c) Suponga que para estimar el número de personas N que pertenecen a un gran consorcio se han seleccionado al azar a 20 de sus trabajadores, a quienes se les ha registrado y colocado un sello en su DNI. Tiempo después, la Dirección de Recursos Humanos tomó un MASs de 100 trabajadores, y encontro que 4 de ellos tenían el sello en el DNI. Por su parte, usted optó más bien por seleccionar secuencialmente al azar y con reemplazamiento trabajadores del consorcio hasta ubicar a 5 con el sello en el DNI, y realizó un total de 127 registros. Obtenga las estimaciones correspondientes de N y sus intervalo de confianza asintóticos al 95 %. Comente.

3. Demuestre que en un MASc la media muestral es el MELI de la media poblacional y que la varianza muestral es una estimador insesgado de σ_N^2.

4. Considere una población finita de tamaño N en la que se desea estudiar una variable estadística y, la cual toma un valor muy pequeño para el primer elemento del marco muestral y_1 y un valor muy grande para el último elemento del marco muestral y_N. Con el propósito de estimar la media de y para esta población, μ, se ha propuesto, sobre la base de un MASs de tamaño n, el estimador:

$$\bar{Y}_c = \begin{cases} \bar{Y} + c & \text{si } y_1 \text{ pertence a la muestra e } y_N \text{ no pertence a la muestra} \\ \bar{Y} - c & \text{si } y_1 \text{ no pertence a la muestra e } y_N \text{ pertence a la muestra} \\ \bar{Y} & \text{en otro caso,} \end{cases}$$

donde c es una constante positiva.
a) ¿Es $\bar{Y}_c$ un estimador insesgado de μ?
b) Halle la varianza de $\bar{Y}_c$.
c) ¿Existen valores de c que hagan que $\bar{Y}_c$, tenga menor varianza que $\bar{Y}$? ¿Contradice esto a que $\bar{Y}$ sea el MELI de μ?

5. Suponga que se desea estimar, con un error no mayor al 3 % y una confianza del 95 %, la prevalencia de una rara enfermedad al interior de una pequeña comunidad de 500 habitantes. Se espera que la proporción de personas de la comunidad que tengan esta enfermedad sea pequeña, lo cual se ha evidenciado en una muestra piloto realizada a 30 de sus habitantes en la que se encontró que solo 2 de ellos tenían la enfermedad.
a) Halle el tamaño de muestra apropiado para este estudio.
b) Puesto que la proporción a estimar es extrema, utilice más bien un IC de Wilson para obtener el tamaño de muestra. Comente la diferencia encontrada con a) e indique cuál de los dos tamaños de muestra utilizaría para el estudio. Justifique.

6. Muestre que en un MASc de tamaño n, sobre una población de tamaño N, el número total de muestras distintas que se podrían tomar es

$$C_n^{N+n-1}.$$

7. En una empresa de 3200 empleados se realizaron dos encuestas independientes por MASs de tamaños 100 y 64 a fin de medir, entre otros, el tiempo que el empleado estima le toma llegar a la empresa or día. Las divisiones de la empresa, que realizaron estas encuestas, no supieron que la otra lo había realizado y al enterarse han decidido unir sus bases de datos.
a) ¿Conforma la media de las 164 observaciones del tiempo de interés un estimador insesgado del tiempo medio de transporte de un empleado a la empresa?
b) Si se define como estimador de la varianza de los tiempos de transporte a

$$S^2 = \frac{100S_1^2 + 64S_2^2}{164},$$

donde S_1^2 y S_2 son las varianzas muestrales de estos tiempos en las encuestas con 100 y 64 empleados, respectivamente, ¿es este un estimador insesgado?
c) Obtenga el error estándar de estimación estimado del estimador en a), si en las muestras de tamaños 100 y 64, se obtuvieron desviaciones estándar muestrales para los tiempos de transporte de 8.625 y 10.162 minutos, respectivamente.

8. Una ciudad cuenta con 720 fábricas, de las cuales 10, 20 y 8 pertenecen, respectivamente, a los consorcios A, B y C. El Ministerio de Trabajo desea hacer un estudio de salud ocupacional en las fábricas de la ciudad. Dado que muchos de los indicadores que se han de estudiar son proporciones, el Ministerio desea tomar un MASs de tamaño n de tal manera que pueda estimar cualquier proporción con un margen de error no mayor a 0.1 y un nivel de confianza del 95 %.
a) ¿Cuál debería ser el tamaño de muestra a tomarse?
b) ¿Con qué probabilidad se seleccionará en la muestra del tamaño tomado en a) a alguna de las fábricas del consorcio B?
c) Suponga que tomada la muestra en a), y dadas las características especiales de los 3 consorcios en mención, el Ministerio ordena que, de ser seleccionada cualquier fábrica de algunos de los consorcios, se seleccione igualmente a todas las fábricas del consorcio elegido. ¿Cuál sería el tamaño de muestra final que esperaría obtener a través de este procedimiento?

9. En cierta área de una ciudad, que contiene 14 848 residencias, se desea estimar el número promedio de personas μ por residencia. Si en un MASs de tamaño 30 se obtuvieron las siguientes cantidades de personas por residencia:

$$5, 6, 3, 3, 2, 3, 3, 3, 4, 4, 3, 2, 7, 4, 3, 5, 4, 4, 3, 3, 4, 3, 3, 1, 2, 4, 3, 4, 2, 4.$$

a) Estime μ y su intervalo de confianza al 95 %.
b) Estime e interprete el CV del número de personas por residencia.
c) Suponga que se desea estimar el número medio anterior con el doble de precisión que la brindada por la muestra anterior. ¿Cuál debería ser el tamaño de muestra para lograr esta precisión?

10. Su distrito, que cuenta con N viviendas, participará en una encuesta por MASs de tamaño n. Suponga que existe una probabilidad constante q de que una vivienda del distrito no responda a esta encuesta. Para prevenir la no respuesta, el supervisor ha decidido, de ser necesario, seleccionar al azar y sin reemplazamiento durante un segundo día un número igual al número de viviendas sin respuesta en el primer día de entre las viviendas aún no seleccionadas.
a) ¿Con qué probabilidad será encuestada su vivienda el primer día?
b) Si en el primer día su vivienda no es seleccionada y no hubo respuestas en M viviendas, ¿con qué probabilidad será seleccionada su vivienda en el segundo día?
c) Si sus padres residen en otra vivienda de su distrito, ¿qué probabilidad existe de que su vivienda y la de sus padres sean seleccionadas?
d) ¿Con qué probabilidad no será posible completar el tamaño de muestra que ha sido planificado para la encuesta?
e) Obtenga d) si $q = 0.06$ y $n = 100$.

11. Para realizar una encuesta de opinión en una población de 150 000 habitantes en la que se encuentran usted y un amigo suyo, se ha diseñado un MASs de tamaño 100.
a) ¿Con qué probabilidad integrará usted la muestra?
b) Suponga ahora que 5 muestras como las anteriores son secuencialmente seleccionadas de esta población mediante un MASs. ¿Qué probabilidad existe de que ni a usted ni a su amigo se les pida su opinión? Asuma que los encuestadores en estas 5 muestras no toman en cuenta el registro de si una persona fue o no seleccionada en otra de las muestras.
c) ¿Con qué probabilidad le pedirán en b) dos veces su opinión?

12. En este capítulo vimos que S^2 es un estimador insesgado de la varianza poblacional σ_N^2 en un MASc y de σ_{N-1}^2 en un MASs, pero ¿qué hay de su varianza?
a) Muestre que

$$S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (Y_i - Y_j)^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i - y_j)^2 \delta_i \delta_j.$$

b) Muestre, usando la fórmula anterior, que S^2 es efectivamente un estimador insesgado.
c) Cho y Cho (2008) han derivado fórmulas para la varianza de S^2, tanto en un esquema MASc como en un MASs. Estas vienen dadas respectivamente por

$$V_{MASc}(S^2) = \frac{1}{n}(\mu_4 - (\frac{n-3}{n-1})\sigma_N^4) \quad y$$

$$V_{MASs}(S^2) = C((Nn - N - n - 1)\mu_4 - (\frac{N^2 n - 3n - 3N^2 + 6N - 3}{N-1})\sigma_N^4),$$

donde: $C = \frac{N(N-n)}{n(n-1)(N-1)(N-2)(N-3)}$ y $\mu_4 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_N)^4$ es el cuarto momento centrado poblacional. Muestre que

$$V_{MASs}(S^2) \to V_{MASc}(S^2), \text{ conforme } N \to \infty.$$

13. Replique el estudio ECE 2019 de la subsección 2.3.2 para la DRE de Lima Metropolitana, pero use, en esta ocasión, los datos de la ECE 2018 para obtener el tamaño de muestra adecuado para el estudio. Además, dado que esta base de datos incluye un indicador de nivel socioeconómico `ISE,` indique mediante un MASs si es que se puede hablar o no de una asociación significativa entre el nivel socioeconómico y los niveles de logro en matemáticas. Use un nivel de significación de $\alpha = 0.05$.

14. Una población cuenta con un total de N personas y es de interés realizar en ella un MASc de tamaño $n = 5$.
a) Halle la función de probabilidad y el valor esperado de la variable aleatoria X que denota el número de personas distintas que contendrá la muestra.
b) Suponga que extraída la muestra anterior es de interés estimar el total τ de una variable y, para lo cual usted multiplicará por una constante C la suma de todos los valores de y en la muestra que correspondan solo a personas distintas. ¿Cuál sería el valor de C que haga de este un estimador insesgado del total?
c) Halle la varianza del estimador construido en b).
d) Utilice los números aleatorios

$$0.327,\ 0.894,\ 0.031,\ 0.289 \text{ y } 0.643,$$

para seleccionar su muestra de una población de 15 personas y reporte el número de personas distintas obtenidas.

15. En un país se ha diseñado una encuesta con el fin de estimar, mediante un MASs, su tasa de desempleo, el cual se cree que cree está en alrededor el 10 % de la PEA (población económicamente activa). En este país, la PEA se define como la población de ciudadanos de 14 años o más de edad y constituye, según el último censo, el 65 % de la población, la cual fue calculada en 2.3 millones de habitantes. Si se quiere estimar la tasa de desempleo con un error no mayor al 1 % y un nivel de confianza del 95 %,
a) ¿Cuál sería el tamaño de muestra a tomar?
b) El costo por cada encuesta se ha estimado en 3 unidades monetarias (u.m.), pero se tiene un presupuesto de tan solo 15 000 u.m. Si se tomará en la muestra la mayor cantidad de personas que pudieran costearse con este presupuesto, ¿cuál sería el margen de error que debería de reportar en este estudio?
c) Suponga que otro interés de la encuesta, es estimar el monto total mensual de ingresos que las personas no desempleadas de la PEA destinan a su consumo. Indique cómo podría estimar este total y su correspondiente error estándar de estimación, ejemplificando esto si en la muestra tomada se encontró una proporción muestral de desocupados del 12.5 % de la PEA, teniendo ellos un gasto promedio de consumos de 4500 u.m. con desviación estádar de 1230 u.m. Sugerencia: De una mirada al siguiente ejercicio.

16. En una zona rural de 3000 viviendas se tomó un MASs de tamaño 100. Un interés del estudio es estimar el consumo total mensual de agua de los hogares que cuentan con servicio de agua y desagüe, τ_d. El problema es que antes de tomarse la muestra no es posible identificar si una vivienda de la zona tiene o no estos servicios.
a) En general, dada una población estadística $\mathcal{P}_y = \{y_1, y_2, \ldots, y_N\}$ y un MASs de ella de tamaño n, muestre que para cierto subconjunto de esta población (dominio d) el estimador

$$\hat{\tau}_d = \frac{N}{n}\sum_{i=1}^{N} y_i\gamma_i\delta_i \quad \text{o} \quad \hat{\tau}_d = \frac{N}{n}\sum_{i=1}^{n} Y_i\gamma_i,$$

donde Y_i es el valor de y para la i-ésima unidad seleccionada en la muestra y γ_i es una variable indicadora (no aleatoria) que vale, respectivamente, 1 o 0 si la i-ésima unidad pertenece o no al dominio d es un estimador insesgado del total τ_d de y para el dominio.
b) Sea la variable y^* que vale y para los elementos del dominio d y 0 en caso contrario, y sea σ^2_{*d} la varianza de $\mathcal{P}_{y^*}$. Si σ^2_d es la varianza de y para los elementos del dominio, muestre que

$$\sigma^2_{*d} = \frac{1}{N-1}((N_d-1)\sigma^2_d + q_d N_d \mu^2_d) \simeq p_d(\sigma^2_d + q_d\mu^2_d),$$

donde N_d es el tamaño del dominio d; μ_d, la media de y en el dominio d; p_d, la proporción de unidades en la población que pertenecen al dominio d y $q_d = 1 - p_d$.
c) Halle la varianza de $\hat{\tau}_d$.
d) Muestre que si se desea estimar τ_d con un error de estimación no mayor a e y una confianza del $100(1-\alpha)\,\%$, el tamaño de muestra apropiado viene dado por

$$n = \frac{((N_d-1)\sigma^2_d + q_d N_d\mu^2_d)z^2_{1-\frac{\alpha}{2}}N^2}{((N_d-1)\sigma^2_d + q_d N_d\mu^2_d)z^2_{1-\frac{\alpha}{2}}N + e^2(N-1)} \simeq \frac{p_d(\sigma^2_d + q_d\mu^2_d)z^2_{1-\frac{\alpha}{2}}N^2}{p_d(\sigma^2_d + q_d N_d\mu^2_d)z^2_{1-\frac{\alpha}{2}}N + e^2}.$$

e) Muestre que el tamaño en d), en caso de que se desee obtener un coeficiente de variación de a lo más CV_0 para el total estimado, se puede aproximar por

$$n = \frac{CV^2_d + q_d}{\frac{CV^2_d + q_d}{N} + p_d CV^2_0},$$

donde

$$CV^2_d = \frac{\sigma^2_d}{\mu^2_d}$$

denota el cuadrado del coeficiente de variación de y en el dominio d.
f) Halle el tamaño de muestra necesario para una encuesta futura que desea estimar τ_d con un margen de error no mayor al millón de litros y una confianza del 95 %. Suponga que en la encuesta actual se encontró que 60 hogares contaban con servicios de agua y desagüe y que en promedio ellos consumieron en el mes 5100 litros con una desviación estándar de 380 litros ¿Qué estimación de τ_d dio la actual encuesta?

17. Consideremos la siguiente base de datos, que llamaremos Province91, tomada del texto de Lehtonen y Pahkinen (2004). Esta contiene información censal de las 32 municipalidades de una de las 14 provincias (Finlandia central) en las que se dividía el país de Finlandia a finales de 1991. En esta se registran para cada municipalidad una variable de estratificación (Stratum con 1 = Urbano y 2 = Rural), de conglomeración (Cluster formado al juntar 4 municipalidades geográficamente vecinas), de población (POP91), de fuerza laboral o población económicamente activa (LAB), del número de personas desempleadas (UE91) y del número de hogares sobre la base del censo de 1985 (HOU85). La base de datos es la siguiente:

Stratum	Cluster	Id	Municipality	POP91	LAB91	UE91	HOU85
1	1	1	Jyväskylä	67200	33786	4123	26881
1	2	2	Jämsä	12907	6016	666	4663
1	2	3	Jämsänkoski	8118	3818	528	3019
1	2	4	Keuruu	12707	5919	760	4896
1	3	5	Saarijärvi	10774	4930	721	3730
1	5	6	Suolahti	6159	3022	457	2389
1	3	7	Äänekoski	11595	5823	767	4264
2	5	8	Hankasalmi	6080	2594	391	2179
2	6	9	Joutsa	4594	2069	194	1823
2	7	10	Jyväskmlk	29349	13727	1623	9230
2	4	11	Kannonkoski	1919	821	153	726
2	4	12	Karstula	5594	2521	341	1868
2	8	13	Kinnula	2324	927	129	675
2	8	14	Kivijärvi	1972	819	128	634
2	3	15	Konginkangas	1636	675	142	556
2	5	16	Konnevesi	3453	1557	201	1215
2	1	17	Korpilahti	5181	2144	239	1793
2	2	18	Kuhmoinen	3357	1448	187	1463
2	4	19	Kyyjärvi	1977	831	94	672
2	5	20	Laukaa	16042	7218	874	4952
2	6	21	Leivonmäki	1370	573	61	545
2	6	22	Luhanka	1153	522	54	435
2	7	23	Multia	2375	1059	119	925
2	1	24	Muurame	6830	3024	296	1853
2	7	25	Petäjävesi	3800	1737	262	1352
2	8	26	Pihtipudas	5654	2543	331	1946
2	4	27	Pylkönmäki	1266	545	98	473
2	3	28	Sumiainen	1426	617	79	485
2	1	29	Säynätsalo	3628	1615	166	1226
2	6	30	Toivakka	2499	1084	127	834
2	7	31	Uurainen	3004	1330	219	932
2	8	32	Viitasaari	8641	4011	568	3119

Usando la librería survey de R, realice tanto un MASc como un MASc de tamaño $n = 8$ para estimar la población total de la provincia y el porcentaje o tasa de desempleo en esta. Reporte en ambos casos los errores estándar de estimación. Compare sus estimaciones con las obtenidas en el texto de Lehtonen y Pahkinen (2004).

18. Usando la base de datos api, obtenga el tamaño de muestra que se requeriría para estimar el índice api del 2000 de tal manera que se tenga para este un CV del 3 % con una confianza del 95 %. Tomada la muestra, estime también el total de matriculados y la proporción de colegios por tipo de escuela. Compare, finalmente, los verdaderos valores (que en un estudio real se desconocen) con las estimaciones encontradas.

19. Mediante un MASs piloto de tamaño n_1 se ha calculado que el tamaño final de muestra a tomarse para estimar la media de una variable y con un máximo error de estimación de e y una confianza del $100(1-\alpha)\,\%$ es n. Un colega sugiere que en vez de seleccionarse las n observaciones bastaría tomarse un MASs de tamaño $n - n_1$ de la población que no ha sido muestreada, pues argumenta que la muestra piloto ya recabó información de y y que juntando esta con la última completaría el tamaño n requerido. ¿Estaría usted de acuerdo con su colega? Justifique.

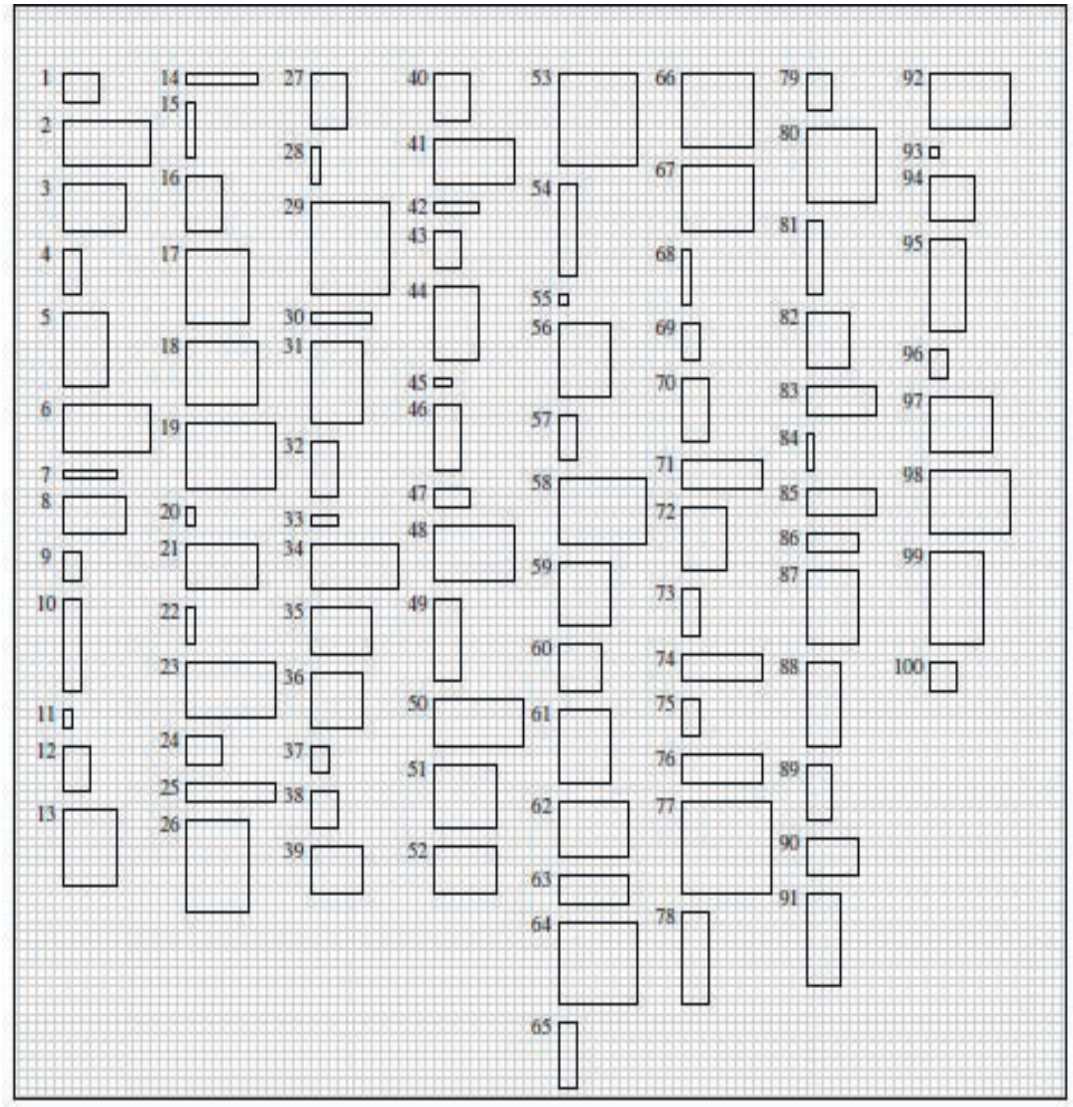
20. Suponga que para un MASs de tamaño n sobre una población de tamaño N se tiene interés en estudiar dos variables estadísticas: x e y.
a) Muestre que la covarianza entre las medias muestrales de estas variables viene dada por

$$Cov(\bar{X}, \bar{Y}) = (1 - \frac{n}{N})\frac{\sigma_{xy}}{n},$$

donde $\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$ es la covarianza poblacional entre x e y y μ_x y μ_y son las medias poblacionales de x e y, respectivamente.
b) Proponga algún estimador insesgado para esta covarianza.

21. La Internet Movie Database (IMDb) es una base de datos en línea que almacena información relacionada con películas, personal de equipo de producción (incluidos directores y productores), actores, series de televisión, programas de televisión, videojuegos, actores de doblaje y, más recientemente, personajes ficticios que aparecen en los medios de entretenimiento visual. Recibe más de 100 millones de usuarios únicos al mes y cuenta con una versión móvil. Una de sus secciones, "The IMDb Top 250", es destinada a ser un listado de las 250 películas con mejor calificación, el cual se basa en calificaciones de los usuarios registrados del sitio web. En esta sección, cada película aparece con una estrella y un ranking de a lo más 10 puntos. Debajo de este ranking uno puede acceder a las calificaciones otorgadas por los usuarios en forma de un histograma. La intención de este miniproyecto es estimar, con un margen de error de a lo más 0.05 puntos y un nivel de confianza del 95 %, la desviación estándar media (como medida de controversia) de los rankings asignados a estas 250 películas.
a) Halle el tamaño de muestra necesario para este estudio.
b) Tome la muestra respectiva y reporte la estimación pedida y con su intervalo de confianza o error estándar de estimación estimado.
c) Según sus resultados, ¿podría decir que *El club de la pelea (1999)* es una película de calificación controversial?

22. En esta actividad sugerida por Gnanadesikan (1997) se tiene la siguiente figura que contiene 100 rectángulos. El objetivo es estimar el área total de todos los rectángulos tomando una muestra de 20 rectángulos, donde se asume que cada cuadradito de la grilla tiene un área de una unidad.



a) Tome un MASs de 20 rectángulos y obtenga un intervalo de confianza al 98 % para el área total.

b) Replique a) pero con un MASc.

c) Compare el intervalo que ha obtenido con el de sus compañeros e indique el porcentaje de estos que contienen la verdadera área que es de 3079 unidades.

23. Luego de realizarse un MASs de tamaño n en una población de tamaño N, se encontró que por error el marco muestral contenía 2 unidades que se repetían, respectivamente, 3 y 7 veces.
a) Halle la probabilidad de selección de cada una de las unidades en la población.
b) Halle la función de probabilidad del número de unidades que deberán descartarse en la encuesta.

24. El sector salud está interesado en saber cuál es la estatura promedio de los habitantes de una región particular que cuenta con 700 habitantes. De los datos de los registros de las clínicas de salud de la región, se realizó un MASs con 35 registros de esta población y se obtuvo:

Obs.	Estatura (mts)	Género	Obs.	Estatura (mts)	Género	Obs.	Estatura (mts)	Género
1	1.65	Hombre	13	1.75	Hombre	25	1.53	Mujer
2	1.80	Hombre	14	1.68	Hombre	26	1.65	Mujer
3	1.84	Hombre	15	1.78	Hombre	27	1.70	Mujer
4	1.83	Hombre	16	1.80	Hombre	28	1.70	Mujer
5	1.73	Hombre	17	1.73	Hombre	29	1.58	Mujer
6	1.83	Hombre	18	1.83	Hombre	30	1.75	Mujer
7	1.80	Hombre	19	1.85	Hombre	31	1.70	Mujer
8	1.85	Hombre	20	1.65	Hombre	32	1.73	Mujer
9	1.80	Hombre	21	1.78	Hombre	33	1.73	Mujer
10	1.78	Hombre	22	1.75	Hombre	34	1.57	Mujer
11	1.85	Hombre	23	1.75	Hombre	35	1.70	Mujer
12	1.80	Hombre	24	1.88	Hombre			

a) Estime la media y varianza de las estaturas en esta población, así como la proporción de mujeres en esta. Puede hacerlo manualmente o con R.
b) ¿Cuál es el error máximo de estimación que se está asumiendo en la estimación de la estatura media para un nivel de confianza del 95 %?
c) Si se hubiese tenido interés en estimar la estatura media de esta población con un margen de error (o error máximo de estimación) de un centímetro a un nivel de confianza del 95 %, ¿hubiese sido suficiente el tamaño de muestra tomado en el estudio?
d) Si en un estudio futuro se desea estimar la estatura media de esta población de tal manera que se tenga un CV no mayor al 0.5 %, ¿cuál sería el tamaño de muestra? ¿Es aquí necesario fijar el nivel de confianza?

25. En una investigación que pretende estudiar caracteriśticas de los colegios y la relación entre la propensión al consumo de alcohol por parte de adolescentes varones del quinto grado de secundaria y variables como el control parental, la regulación emocional y la madurez social, se desea tomar un MASs de colegios con alumnos del quinto grado de secundaria de la UGEL 03. Puesto que la propensión se medirá a nivel de colegios mediante una proporción, es de interés estimar esta proporción con un margen de error no mayor a 0.06 y un nivel de confianza del 95 %. Usando en lo posible el paquete survey de R,
a) Halle el tamaño de muestra requerido para este estudio. Para su marco muestral puede usar la siguiente página web del Ministerio de Educación:

`http://escale.minedu.gob.pe/web/inicio/padron-de-iiee ,`

la cual contiene información de todos los colegios del país basada en el último censo nacional escolar b) Tome la muestra anterior y estime basándose en ella el total actual de alumnos matriculados en la UGEL 03, así como la proporción de colegios de gestión privada en esta UGEL. En ambos casos obtenga el correspondiente error de estimación estimado.

26. En el conteo rápido de votos realizado a 1600 urnas seleccionadas al azar de una gran población se obtuvo que 812 votaron por el candidato opositor, 480 lo hicieron por el candidato de gobierno, 50 votaron en blanco y el resto fueron votos inválidos. Al 95 % de confianza,
a) ¿Cuál es el el máximo error de estimación que se comete en esta encuesta al estimar la proporción de ciudadanos que votan por el candidato opositor?
b) Mediante un intervalo de confianza, ¿podría afirmar que el candidato opositor ganará las elecciones? Para esto se requiere el 50 % de votos válidos más uno.

27. Suponga que es de su interés estimar el tiempo medio que una persona se tardaría en llegar desde el campus de la PUCP al centro comercial Real Plaza Salaverry en auto. Una manera directa de medir este tiempo es a través del aplicativo Google Maps, el cual se puede descargar gratuitamente en cualquier PC, laptop o celular. Este aplicativo calcula, por medio del GPS, el tiempo que una persona se demoraría en llegar de un lugar a otro bajo distintos medios de transporte. Estos tiempos, sin embargo, cambian según el horario, en especial si el medio es un auto, debido a congestiones en el tráfico, accidentes u otros. El aplicativo también brinda varias rutas alternativas, de las cuales usted deberá tomar la de menor tiempo. En este problema se le pide estimar el tiempo medio anterior y su intervalo de confianza al 95 % mediante un MASs de tal manera que su error de estimación sea de a lo más de un minuto. Para su procedimiento de selección (ignorando aspectos estacionales) divida una semana completa de 7 días en 336 períodos de media hora cada uno. Tome luego al azar y sin reemplazamiento el número de períodos adecuados y en cada período seleccionado registre en cualquier momento de ese período la medición del tiempo en minutos dada por el aplicativo. Reporte, finalmente, la estimación del tiempo medio y del intervalo de confianza y compruebe si el error máximo predeterminado es el especificado.

28. En la subsección 2.3.1 obtuvimos el error estándar de estimación para la diferencia de medias del índice de rendimiento api para 1999 y el 2000.
a) Tome en esta base de datos un MASs de tamaño $n = 500$ y estime con la librería survey la diferencia de medias del índice api para estos años.
b) Obtenga, con la librería survey, un intervalo de confianza al 95 % para la diferencia anterior.
c) Con la misma muestra tomada en a) obtenga el intervalo de confianza b), pero ahora sin usar el paquete survey.

Capítulo 3

Muestreo aleatorio estratificado

3.1. Introduccción

Cuando la variable de interés asume en promedio distintos valores sobre diferentes subconjuntos de la población, uno podría obtener estimaciones mucho más precisas de tomar en cuenta esta segmentación. En una muestra estratificada, la población se particiona en H subconjuntos o estratos que tienen la propiedad de ser heterogéneos entre sí pero homogéneos al interior. La idea aquí es extraer una muestra independiente en cada estrato (usualmente mediante un MASs) y, posteriormente, reunir esta información para obtener estimaciones globales de la población.

Entre las razones para optar por un muestreo aleatorio estratificado podemos citar las siguientes:

- Queremos protegernos contra la posibilidad de obtener un mala muestra, en el sentido de que algún estrato no esté o esté pobremente representado.
- Es probable que queramos datos de precisión conocida sobre cada estrato.
- La muestra estratificada podría administrarse más convenientemente, a un costo menor, reduciendo el tamaño de muestra en los estratos más caros e incrementando este tamaño en lo más baratos.
- El muestreo estratificado dará, si se hace correctamente, estimaciones más precisas para toda la población.

3.2. Teoría del muestreo aleatorio estratificado

Supongamos que una población de N unidades está particionada en H estratos, donde cada estrato h posee N_h unidades ($N_1 + N_2 + \ldots + N_H = N$). En el muestreo aleatorio

estratificado, que simplemente lo llamaremos MAE, seleccionaremos en forma independiente muestras aleatorias simples de tamaño n_h para cada estrato h (específicamente mediante un MASs [1]). Así, tendremos que

$$n = n_1 + n_2 + \ldots + n_H$$

representará el tamaño de muestra en la población y se obtendrán los siguientes parámetros y estimadores puntuales de interés, donde y_{hi} denotará el valor de la variable estadística de interés y en el i-ésimo objeto del estrato h y δ_{hi} denotará, como antes, la variable aleatoria dicotómica que vale 1 si el i-ésimo elemento del estrato h es seleccionado en la muestra de tamaño n_h o 0 en caso contrario.

Denominación	Parámetro poblacional	Estimador puntual
Media en el estrato h	$\mu_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$	$\bar{Y}_h = \frac{1}{n_h}\sum_{i=1}^{N_h} y_{hi}\delta_{hi}$
Varianza en el estrato h	$\sigma_h^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{hi}-\mu_h)^2$	$S_h^2 = \frac{1}{n_h-1}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y}_h)^2\delta_{hi}$
Media global	$\mu = \sum_{h=1}^{H}\frac{N_h}{N}\mu_h$	$\bar{Y} = \sum_{h=1}^{H}\frac{N_h}{N}\bar{Y}_h$

Observe que la tabla anterior incluye indirectamente el caso de la proporción, de tomarse y como una variable dicotómica.

Veamos ahora las propiedades de los estimadores puntuales en el lado derecho de la tabla. Para ello recordemos que en cada estrato se ha tomado un MASs y, por tanto, $\bar{Y}_h$ y S_h^2 son estimadores insesgados de μ_h y σ_h^2, respectivamente. Más aún, se tiene que

$$E(\bar{Y}) = \sum_{h=1}^{H}\frac{N_h}{N}E(\bar{Y}_h) = \sum_{h=1}^{H}\frac{N_h}{N}\mu_h = \mu$$

y, por tanto, $\bar{Y}$ es también un estimador insesgado de μ. La varianza de este estimador está dada por

$$V(\bar{Y}) = \sum_{h=1}^{H}(\frac{N_h}{N})^2 V(\bar{Y}_h) = \sum_{h=1}^{H}(\frac{N_h}{N})^2(1-\frac{n_h}{N_h})\frac{\sigma_h^2}{n_h}.$$

Por tanto, un estimador insesgado natural de esta varianza se obtiene al reemplazar la varianza poblacional en el estrato h por su varianza muestral S_h^2, dando así lugar al estimador insesgado

$$\hat{V}(\bar{Y}) = \sum_{h=1}^{H}(\frac{N_h}{N})^2 V(\bar{Y}_h) = \sum_{h=1}^{H}(\frac{N_h}{N})^2(1-\frac{n_h}{N_h})\frac{S_h^2}{n_h} \tag{3.1}$$

y al error estándar de estimación de $\bar{Y}$ estimado:

$$\hat{SE}(\bar{Y}) = \sqrt{\hat{V}(\bar{Y})} = \sqrt{\sum_{h=1}^{H}(\frac{N_h}{N})^2(1-\frac{n_h}{N_h})\frac{S_h^2}{n_h}}.$$

[1] Podríamos tomar también un MASc, que es más simple, pero poco utilizado en la práctica. Los resultados aquí desarrollados se pueden también aplicar a este último esquema con las modificaciones mínimas derivadas de no incluir el factor de corrección para poblaciones finitas.

Por el TLC es posible deducir que un intervalo de confianza aproximado al $100(1-\alpha)\,\%$ para μ, cuando los tamaños de muestra en cada estrato o la cantidad de estratos es grande, viene dado por

$$[\bar{Y} - z_{1-\frac{\alpha}{2}}\hat{SE}(\bar{Y}), \bar{Y} + z_{1-\frac{\alpha}{2}}\hat{SE}(\bar{Y})]$$

Esta aproximación puede no ser adecuada si los tamaños de muestra en los estratos son pequeños, en cuyo caso se recomienda el uso de una aproximación t dada por

$$[\bar{Y} - t_{1-\frac{\alpha}{2}}(d)\hat{SE}(\bar{Y}), \bar{Y} + t_{1-\frac{\alpha}{2}}(d)\hat{SE}(\bar{Y})],$$

donde los grados de libertad pueden obtenerse de la aproximación de Satterthwaite (1946) por

$$d = \frac{(\sum_{h=1}^{H} c_h S_h^2)^2}{\sum_{h=1}^{H} \frac{1}{n_h - 1}(c_h S_h^2)^2},$$

siendo $c_h = \frac{N_h(N_h - n_h)}{n_h}$.

En el caso de la estimación de una proporción basta considerar en el desarrollo anterior a una variable dicotómica y, así el error estándar de estimación estimado para la proporción global estimada

$$\bar{p} = \sum_{h=1}^{H} \frac{N_h}{N}\bar{p}_h,$$

donde $\bar{p}_h$ es la proporción muestral en el estrato h, se reduce a

$$\hat{SE}(\bar{p}) = \sqrt{\sum_{h=1}^{H}(\frac{N_h}{N})^2(1 - \frac{n_h}{N_h})\frac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}}.$$

3.3. Pesos de muestreo y efectos de diseño

En todo el desarrollo anterior hemos utilizado como notación las variables δ_{hi}. Estas determinan la aleatoriedad de los estimadores y son variables indicadoras de si un elemento en la población es o no seleccionado en la muestra del estrato h. Así, la probabilidad de elegir la unidad i en el estrato h viene dada, por ser este un MASs, por

$$P(\delta_{hi} = 1) = \frac{n_h}{N_h}.$$

Otra manera equivalente de representar una muestra para el estrato h, prescindiendo de las variables δ_{hi}, es, como expresamos en el MAS, mediante una colección de variables aleatorias $Y_{h1}, Y_{h2}, \ldots, Y_{hn_h}$ que denotan los valores de la variable estadística y que se obtendrán secuencialmente en cada selección del estrato h. Si adoptamos esta notación, podríamos reescribir la media muestral de un MAE como

$$\bar{Y} = \sum_{h=1}^{H} \frac{N_h}{N}\bar{Y}_h = \sum_{h=1}^{H} \frac{N_h}{N}(\frac{1}{n_h}\sum_{i=1}^{n_h} Y_{hi})$$

o como

$$\bar{Y} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \omega_{hi} Y_{hi},$$

donde ω_{hi} es un peso o factor de expansión dado por

$$\omega_{hi} = \frac{N_h}{n_h} = \frac{1}{P(\delta_{hi} = 1)}.$$

Este se puede interpretar como el número de unidades en la población del estrato h que es representada por cada miembro de la muestra. Si, por ejemplo, la población tiene 2000 sujetos y ella está estratificada por sexo con 1200 hombres y 800 mujeres, entonces en un MAE de 200 hombres y 200 mujeres, cada hombre de la muestra tiene un peso de 6 y cada mujer un peso de 4. En otras palabras, cada hombre se representa a sí mismo y a 5 más que no están en la muestra; mientras que cada mujer se representa a sí misma y a 3 más que no están en la muestra. Luego, como cada unidad de la muestra se puede pensar que representa a cierta cantidad de elementos de la población, la muestra completa puede pensarse que representa a toda la población. De aquí el nombre alternativo para los ω_{hi} de factores de expansión.

Otro elemento importante a lo largo del curso lo constituirán los efectos de diseño. El efecto de un diseño al estimar un parámetro θ mediante un estimador $\hat{\theta}$ se define como el parámetro $deff$, que resulta del cociente entre la varianza de $\hat{\theta}$ bajo el diseño en estudio y la varianza de $\hat{\theta}$ bajo un MASs, ambos con un mismo tamaño de muestra. El diseño MASs en el denominador es tomado aquí como un diseño de referencia o *benchmark*. Formalmente se expresa así:

$$deff(\hat{\theta}) = \frac{V_c(\hat{\theta})}{V_{MASs}(\hat{\theta})}.$$

Un diseño, por tanto, será más eficiente mientras su $deff$ sea cada cada vez menor que 1. Ello, en la práctica, como más adelante veremos en el muestreo complejo, es difícil de alcanzar. El efecto de diseño de un MASc en la estimación de la media es, por ejemplo:

$$deff(\bar{Y}) = \frac{V_{MASc}(\bar{Y})}{V_{MASs}(\bar{Y})} = \frac{\sigma_N^2/n}{(1 - \frac{n}{N})\sigma_{N-1}^2/n} = \frac{N-1}{N-n}.$$

Así, un MASc resulta ser más ineficiente que un MASs al estimar la media; aunque para tamaños de población N suficientemente grandes, tal pérdida de eficiencia es mínima. Note en este caso que el efecto de diseño se obtiene de forma directa, lo cual en general no es cierto, pues tal dependerá de algunos parámetros poblacionales, los cuales requieren estimarse. El problema con la estimación del $deff$ es que para hacerlo solo contamos con la data del diseño utilizado y no con la data bajo el MASs. En tal sentido, se debe ver cómo estimar $V_{MASs}(\hat{\theta})$ con la data proveniente del diseño complejo. Una manera de hacer esto en el MAE se muestra en el ejercicio 3.7.8.

En R, y particularmente en el paquete `survey`,el cálculo de las estimaciones de los efectos de diseño se encuentra disponible de solicitarse la opción `deff`= T. En el caso de estimarse la media bajo un MAE, este nos provee de una estimación de

$$\hat{deff} = \frac{\hat{V}_{MAE}(\bar{Y})}{\hat{V}_{MASs}(\bar{Y})},$$

donde $\hat{V}_{MAE}(\bar{Y})$ se calcula mediante (3.1) y $\hat{V}_{MASs}(\bar{Y})$ por

$$\hat{V}_{MASs}(\bar{Y}) = (1 - \frac{n}{N})\frac{\hat{\sigma}^2}{n},$$

siendo $\hat{\sigma}^2$ una estimación de la varianza de la población de la variable de interés y bajo un MAE, obtenido a través del objeto diseño diseMAE (definido por `svydesign`). Esta estimación, valga aclarar, no es la misma a la dada por la del ejercicio 3.7.8, ya que utiliza, como lo veremos en el capítulo 5, los pesos de muestreo del diseño (sea este un MAE o no) y una función de probabilidad empírica ponderada derivada de tales pesos.

En un MAE, los efectos de diseño tienden por lo general a ser menores que 1 e indican la mayor eficiencia de un MAE con respecto a un MASs, sobre todo si la variable de estratificación logra bien separar a los estratos en grupos relativamente homogéneos. Una ilustración del cálculo de estos efectos y de sus estimaciones se presenta en el siguiente ejemplo.

Ejemplo 3.1. *Considere una población de $N = 8$ domicilios, donde son conocidas las variables renta familiar en unidades monetarias y estrato socioeconómico (A = alto o B = bajo). Los valores de estas variables se resumen en la siguiente tabla:*

Unidad	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Renta	*13*	*17*	*6*	*5*	*10*	*12*	*19*	*6*
Estrato	*B*	*A*	*B*	*B*	*B*	*A*	*A*	*B*

A fin de estimar la renta media familiar, se decide efectuar un MAE con 2 observaciones por estrato. Obtenga el efecto de diseño de este MAE y estímelo en caso se desconozcan los valores de toda la población.

Solución: Calculemos primero el efecto de diseño a través del código siguiente:

```
N = 8
n = 4
Nh = c(3,5)
nh = c(2,2)
Renta = c(13,17,6,5,10,12,19,6)
NSE = c("B","A","B","B","B","A","A","B")
RR = data.frame(Renta,NSE)
Vmae = sum((Nh/N)^2*(1-nh/Nh)*by(Renta,NSE,var)/nh)
```

```
Vmas = (1-n/N)*var(Renta)/n
(deff = Vmae/Vmas)

## [1] 0.482
```

El MAE es, en este caso, mucho más eficiente que un MASs con un efecto de diseño marcadamente menor que 1, lo cual se preveía dadas las marcadas diferencias de rentas entre ambos estratos socioeconómicos. Tal efecto, sin embargo, se ha podido calcular solo porque contamos con la data poblacional. En la mayoría de situaciones, esta es desconocida y requerirá ser estimada con los valores de la muestra. Si realizamos el MAE pedido con

```
set.seed(12345)
RRm = RR[c(sample(which(RR$NSE=="A"),2),sample(which(RR$NSE=="B"),2)),]
```

una manera de estimar el efecto de diseño podría ser utilizando lo presentado en el ejercicio 3.7.8. Bajo este enfoque obtendremos la siguiente estimación:

```
Vmae_e = sum((Nh/N)^2*(1-nh/Nh)*by(RRm$Renta,RRm$NSE,var)/nh)
Ybar = sum((Nh/N)*by(RRm$Renta,RRm$NSE,mean))
s2y = by(RRm$Renta,RRm$NSE,function(x) sum(x^2))
VYbarmas_e = (N-n)/(n*(N-1))*(sum((Nh/(nh*N))*s2y) - Ybar^2 + Vmae_e)
(deff1_e = Vmae_e/VYbarmas_e)

## [1] 0.486
```

La estimación brindada por R, por otro lado, la podremos calcular por

```
diseMAE = svydesign(ids=~1,strata= ~NSE,fpc=c(rep(3,2),rep(5,2)),data=RRm)
(deff2_e = Vmae_e/((1-n/N)*coef(svyvar(~Renta,diseMAE))/n))

## Renta
## 0.447
```

o mucho más directamente mediante

```
svymean(~Renta,diseMAE,deff=T)

##        mean    SE DEff
## Renta 10.81  1.23 0.45
```

□

3.4. Tamaños de muestra

El cálculo de tamaños de muestra en un MAE involucra no solo saber cuántas unidades n seleccionar en la población, sino también cuántas unidades n_h en cada estrato. Para ello requeriremos fijar algún criterio. Dos son los criterios más utilizados. El primero sigue la línea de lo que vimos en el MAS; es decir, garantizar como máximo un error de estimación predeterminado e en la estimación buscada, pero ahora buscando minimizar los costos de muestreo. El segundo consiste en minimizar el error en la estimación bajo un presupuesto C fijo. En ambas situaciones, se asumirá que el presupuesto C tiene una estructura lineal; es decir, es de la forma $C = c_0 + \sum_{h=1}^{H} n_h c_h$, donde c_0 es un costo fijo de muestreo y c_h un costo variable por unidad seleccionada, el cual depende del estrato. Para explicitar los tamaños de muestra supongamos ahora que nuestro interés recae en estimar la media global μ. Dado que el error de estimación depende monotónamente de la varianza de este estimador y que para minimizar los costos requerimos solo minimizar los costos variables, nuestro problema se reducirá, bajo estos criterios, a encontrar los tamaños de muestra por estrato n_h que minimicen la varianza estimada del estimador $\bar{Y}$ de μ:

$$\hat{V}_{est} = \sum_{h=1}^{H} (\frac{N_h}{N})^2 (1 - \frac{n_h}{N_h}) \frac{S_h^2}{n_h}$$

o equivalentemente:

$$\sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{S_h^2}{n_h}$$

o el costo total de muestreo:

$$C = c_0 + \sum_{h=1}^{H} c_h n_h$$

o equivalentemente:

$$\tilde{C} = C - c_0 = \sum_{h=1}^{H} c_h n_h,$$

sujeto a fijar uno de ellos. Dado que los tamaños de muestra n_h por estrato serán finalmente una fracción a_h de n; vale decir

$$n_h = a_h n,$$

la proposición siguiente nos brinda la distribución óptima de los a_h que resuelve el problema de minimización dual anterior.

Proposición 3.1. *En un MAE, con función de costo lineal, la varianza $\hat{V}_{est}$ es mínima para un costo total fijo o este costo es mínimo para una varianza $\hat{V}_{est}$ fija si*

$$a_h = \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{j=1}^{H} \frac{N_j S_j}{\sqrt{c_j}}}. \tag{3.2}$$

Demostración: *Probaremos este resultado basándonos en la celebrada desigualdad de Cauchy-Schwartz. Recordemos que esta nos dice que el valor absoluto del producto interno de dos vectores es siempre menor o igual que el producto de sus normas. En* $\mathbb{R}^n$ *esto se traduce como sigue: si* $a_1, a_2, \ldots, a_n$ *y* $b_1, b_2, \ldots, b_n$ *son números reales cualesquieras, entonces*

$$(\sum_{i=1}^{n} a_i b_i)^2 \leq (\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2).$$

Note que esta desigualdad se convierte en una igualdad si para todo $i = 1, 2, \ldots, n$*:*

$$\frac{a_i}{b_i} = k, \tag{3.3}$$

siendo k *una constante. Ello significa, en otras palabras, que la función*

$$f(a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n) = (\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)$$

se minimiza si se cumple (3.3).

La asignación óptima se obtendrá consecuentemente al minimizar, con respecto a los n_h*, el producto*

$$(\sum_{h=1}^{H} c_h n_h)(\sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{S_h^2}{n_h}),$$

sujeto a que uno de los dos términos en este producto sea fijo. Por lo previamente desarrollado, tenemos entonces que la solución de este problema se obtendrá cuando para cada estrato h *se cumpla*

$$\frac{\sqrt{c_h n_h}}{(\frac{N_h}{N})\frac{S_h}{\sqrt{n_h}}} = k,$$

donde k *es una constante. Despejando obtenemos que*

$$n_h = k \frac{N_h S_h}{N\sqrt{c_h}}.$$

Más aún, como $n = \sum_{h=1}^{n} n_h$*, se tiene que* $k = \frac{n}{\sum_{j=1}^{H} \frac{N_j S_j}{N\sqrt{c_j}}}$*. Esto nos conduce a los* a_h *descritos en la proposición.* ■

Dependiendo del propósito de la encuesta, el tamaño de muestra total n se obtendrá en el caso de la minimización de los costos como el valor n que resuelva

$$e = z_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{H} (1 - \frac{a_h n}{N_h})(\frac{N_h}{N})^2 \frac{S_h^2}{a_h n}}$$

para un error máximo de estimación de la media prefijado e, o más explícitamente mediante

$$n = \frac{\sum_{h=1}^{H} \frac{N_h^2}{a_h} S_h^2}{(\frac{Ne}{z_{1-\frac{\alpha}{2}}})^2 + \sum_{h=1}^{H} N_h S_h^2}.$$

Por otro lado, si el propósito es minimizar el error en la estimación o equivalentemente la varianza del estimador, bajo un presupuesto C fijo, el tamaño total de muestra n vendrá dado por

$$n = \frac{C - c_0}{\sum_{h=1}^{H} c_h a_h}.$$

Naturalmente, todo esto funciona también para el caso de la proporción, con la única modificación que las varianzas muestrales S_h^2 vienen dadas en este caso por $S_h^2 = \frac{N_h \bar{p}_h (1-\bar{p}_h)}{N_h - 1}$, siendo $\bar{p}_h$ la proporción de éxitos muestral en el estrato h.

De la proposición anterior se desprenden los siguientes casos particulares:

- **Asignación proporcional.** Se da cuando

$$a_h = \frac{N_h}{N} \ ;$$

  es decir, cuando los tamaños de muestra en cada estrato se toman proporcionalmente al tamaño del estrato. Este es un caso particular de (3.2) si se consideran varianzas y costos iguales.

- **Asignación de Neyman.** Se da cuando

$$a_h = \frac{N_h S_h}{\sum_{j=1}^{H} N_j S_j} \ ;$$

  es decir, cuando los tamaños de muestra en cada estrato se toman de manera directamente proporcional a la variabilidad y al tamaño del estrato. Este es un caso particular de (3.2) si se consideran costos iguales.

Observación: Si bien en la selección del tamaño de muestra hemos, hasta el momento, buscado controlar el error de estimación de la media o proporción global, tal estrategia no es única. En muchos estudios resulta más conveniente, para obtener mayor precisión en los estratos, determinar el tamaño global n del estudio como un agregado de los tamaños de muestra por estrato, donde estos se calculan mediante un MASs al fijarse los errores de estimación máximo por estrato. Esta técnica está relacionada con la idea de dominios, tema que discutiremos luego del siguiente ejemplo.

Ejemplo 3.2. *En el siguiente ejemplo, tomado de Mendenhall et al. (2007), una empresa publicitaria tiene interés en determinar cómo enfatizar la publicidad televisiva en una determinada región, y decide realizar un muestreo aleatorio estratificado para estimar el número*

promedio de horas por semana que se ve televisión en los hogares de la región. Esta comprende de dos pueblos, A y B, y un área rural, los cuales serán tomados como estratos. El pueblo A está en torno a una fábrica, y la mayoría de los hogares son de trabajadores industriales con niños en edad escolar. El pueblo B es un suburbio exclusivo de una ciudad vecina y consta de habitantes mayores con pocos niños en casa. Existen 155 hogares en el pueblo A, 62 en el pueblo B y 93 en el área rural. Puesto que la información se recopilará mediante encuesta con visita a los hogares, la empresa debe de tomar en cuenta el costo de una observación. El costo por observación en cada pueblo se ha estimado en 9 dólares y en 16 dólares para el área rural debido a costos de transporte. Si las desviaciones estándar del número de horas que se ve televisión (aproximadas por las varianzas muestrales de una encuesta previa) son de 5, 15 y 10, respectivamente, para el pueblo A, B y área rural, encuentre el tamaño global n y los tamaños de muestra por estrato que permitan a la empresa estimar, con el mínimo costo, el tiempo medio que se ve televisión con un límite para el error de estimación de una hora y un nivel de confianza del 95 %.

Solución: Según los datos, tenemos la siguiente tabla para los tamaños de muestra por estrato (N_h), costos unitarios de muestreo por estrato (c_h), desviaciones estándar estimadas por estrato (S_h) y, consecuentemente, asignaciones óptimas por estrato (a_h):

Estrato (h)	N_h	c_h	S_h	$\frac{N_h S_h}{\sqrt{c_h}}$	a_h
Pueblo A	*156*	*9*	*5*	*258.33333*	*0.32258*
Pueblo B	*62*	*9*	*15*	*310*	*0.3871*
Área rural	*93*	*16*	*10*	*232.5*	*0.29032*
Suma				*800.83333*	

Puesto que la intención en este estudio es obtener un error de estimación de a lo más una hora ($e = 1$) con un nivel de confianza del 95 % y un mínimo costo, el tamaño de muestra del estudio estará dado por

$$n = \frac{\sum_{h=1}^{H} \frac{N_h^2}{a_h} S_h^2}{(\frac{311}{1.96})^2 + \sum_{h=1}^{H} N_h S_h^2} = 135.6977 \simeq 136.$$

Deberemos, finalmente, distribuir estas encuestas a tomar en los estratos, obteniéndose así, los siguientes tamaños por estrato para, respectivamente, los pueblos A, B y el área rural:

$$n_1 = 0.32258 \times 136 = 43.87088 \simeq 44,$$

$$n_2 = 0.3871 \times 136 = 52.6456 \simeq 53$$

y $n_3 = 0.29032 \times 136 = 39.48352 \simeq 39.$ □

3.5. Dominios

Un dominio, según Kish (1965), se refiere a una subdivisión de la población para el cual se ha planificado, fijado y seleccionado una muestra a fin de proporcionar resultados específicos para ella bajo un conocido margen de error. Ello significa que los resultados para cada dominio se proporcionan con una precisión determinada y que el resultado global se deriva de una combinación adecuada de los resultados de todos los dominios. Por lo común, los dominios coinciden con ciertas unidades político-administrativas, tales como regiones, provincias, distritos, comunidades, etc. aunque también pueden ser el producto del cruce de dos o más variables de interés en la población. De considerarse el uso de dominios, el muestreo en términos prácticos se realiza como si cada uno de estos fuese una población de la que hay que extraer una muestra representativa. Naturalmente, tiene que ocurrir que la muestra agregada de todos los dominios a su vez cumpla los requisitos de representatividad que se impongan a la población global. Dado que se requiere tener control sobre cada dominio, una pregunta natural es entonces por qué no hacer que cada dominio represente un estrato. Si bien ello es posible, podría, como suele ocurrir, que el marco muestral no nos provea de información sobre la membresía de cada objeto al dominio, ya que esta información podría no conocerse sino quizás hasta después de culminado el muestreo. Otra razón para no usar dominios como estratos es que ello puede resultar poco práctico cuando estos son disjuntos y definidos, por ejemplo, a través del cruce de dos o más variables categóricas, las cuales por la cantidad de sus modalidades podrían generar demasiados estratos. Adicionalmente, cabe tener en cuenta que la idea de los dominios difiere de la de los estratos en el sentido que no es necesario que los primeros conformen necesariamente grupos heterogénos entre sí y homogéneos al interior; pues estos no se diseñan con este fin, sino con la finalidad de conocer más lo que ocurra al interior de cada dominio y cómo se compara este con otros dominios. Un ejemplo interesante del uso de dominios por MASs fue presentado en el ejercicio 14 del capítulo anterior. En ese ejemplo asumimos que el número de viviendas en el dominio d de los que contaban con servicio de agua y desag$\acute{}$ue, N_d, no necesariamente se podía conocer. En general, como aquí, N_d requerirá por lo común estimarse, lo cual introducirá una variabilidad extra y complejizará los cálculos. Imaginemos, por ejemplo, una encuesta de viviendas en las que los dominios de interés esten centrados en ciertas minorías a los cuales se les desea hacer cierta intervención. Uno de estos dominios podría ser, por ejemplo, el de mujeres quechuahablantes. Claramente, aquí solo será posible conocer esta condición en la entrevista, más no previamente en el marco muestral de viviendas y, por tanto, uno desconocerá el número de mujeres quechuahablantes en la población. Aun cuando, como se hace en la práctica, dicha cantidad se reemplace por su número esperado o estimado, se sugiere que el tamaño de muestral obtenido para el dominio bajo esta aproximación se incremente ligeramente a fin de cubrir la variabilidad extra descartada. Para formalizar esto veamos el siguiente desarrollo.

Supongamos que deseamos estimar la media de una variable estadística y para un dominio d bajo un MAE. Para esto podríamos usar un estimador de razón combinado que estime tanto el total τ_d en el dominio como su número de unidades N_d; es decir,

$$\bar{Y}_d = \frac{\hat{\tau}_d}{\hat{N}_d} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\omega_{hi}Y_{hi}\gamma_{di|h}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\omega_{hi}\gamma_{di|h}} = \frac{\sum_{h=1}^{H}\hat{\tau}_{dh}}{\sum_{h=1}^{H}N_h\bar{p}_{dh}},$$

siendo $\omega_{hi} = \frac{N_h}{n_h}$ un peso o factor de expansión; $\gamma_{di|h}$, un indicador no aleatorio 0-1 que vale 1 si la i-ésima unidad seleccionada en el estrato h pertenece al dominio d; $\bar{p}_{dh} = \frac{n_{dh}}{n_h}$, la proporción muestral de unidades en el estrato h que pertenecen al dominio d y $\hat{\tau}_{dh}$ el estimador del total de y para el dominio d del estrato h, el cual describimos en el ejercicio 14 del capítulo 2. Puesto que este es un estimador de razón combinado, como el que se estudiará en el capítulo 5, se sigue de (5.7) y de la parte b) del ejercicio 14 en el capítulo 2 que una aproximación de la varianza de $\bar{Y}_d$ viene dada por

$$V(\bar{Y}_d) = \frac{1}{N_d^2}\sum_{h=1}^{H}N_h^2(1-\frac{n_h}{N_h})\frac{\sigma_{*hz}^2}{n_h}$$

$$= \frac{1}{N_d^2}\sum_{h=1}^{H}\frac{N_h^2}{n_h}(1-\frac{n_h}{N_h})\left((\frac{N_{dh}-1}{N_h-1})\sigma_{dh}^2 + \frac{N_{dh}}{N_h-1}(1-\frac{N_{dh}}{N_h})(\mu_{dh}-\mu_d)^2\right), \qquad (3.4)$$

donde σ_{*hz}^2 es la varianza de todos los valores $z_{*ih} = (y_{ih} - \frac{\tau_d}{N_d})\gamma_{id|h}$ en el estrato h, μ_d la media de todas las N_d unidades del dominio d, N_{dh} el número de unidades en el estrato h que pertenecen al dominio d y μ_{dh} y σ_{dh}^2 la media y varianza de estas últimas unidades.

Una simplificación de (3.4) puede obtenerse si asumimos que la proporción muestral $\bar{p}_{dh}$ es más o menos la misma que la proporción poblacional respectiva $p_{dh} = \frac{N_{dh}}{N_h}$. Ello nos lleva a la aproximación

$$V(\bar{Y}_d) = \sum_{h=1}^{H}(\frac{N_{dh}}{N_d})^2\frac{1}{n_h}(1-\frac{n_h}{N_h})\left(\sigma_{dh}^2 + q_{dh}(\mu_d h - \mu_d)^2\right),$$

donde $q_{dh} = 1 - p_{dh}$. Para la consideración de los tamaños de muestra, podríamos fijar una asignación a los dominios del número de unidades para la muestra del estrato h igual a $n_{dh} = n_h p_{dh}$. Así, si substituimos ello en la ecuación anterior, obtendremos que

$$V(\bar{Y}_d) = \sum_{h=1}^{H}(\frac{N_h}{N})^2(1-\frac{n_h}{N_h})p_{dh}(\frac{N}{N_d})^2\frac{(\sigma_{dh}^2 + q_{dh}(\mu_{dh}-\mu_d)^2)}{n_h}.$$

Consecuentemente, podríamos usar los métodos de asignación estudiados en la sección 3.4, luego de reemplazar S_h^2 por una estimación de $\sigma_{*dh}^2 = p_{dh}(\frac{N}{N_d})^2\left(\sigma_{dh}^2 + q_{dh}(\mu_{dh}-\mu_d)^2\right)$.

Por otro lado, el estimador natural para la varianza (3.4) de la media en el dominio d será

$$\hat{V}(\bar{Y}_d) = \frac{1}{\hat{N}_d^2}\sum_{h=1}^{H}N_h^2(1-\frac{n_h}{N_h})\frac{\hat{\sigma}_{*hz}^2}{n_h}, \qquad (3.5)$$

donde $\hat{\sigma}^2_{*hz}$ denota la varianza muestral de los valores $z_{*hi} = (y_{hi} - \hat{\theta})\gamma_{di|h}$ en el estrato h. Si bien los cálculos parecen complicados, tenemos por fortuna que estos se encuentran implementados en el paquete survey de R a través del comando `svyby`. Este nos provee, por ejemplo, de las estimaciones de las medias por dominio $\bar{Y}_d$ y sus errores estándar de estimación estimados, los que se obtienen como la raíz cuadrada de (3.5). En la siguiente sección ilustraremos el uso de tal comando.

3.6. Uso del paquete survey

3.6.1. MAE con la base de datos api

Para seguir capacitándonos en el uso del paquete survey de R, retomemos el análisis de la base de datos api y supongamos que ahora estamos interesados en un MAE de tamaño 200, donde como criterio de estratificación usaremos el tipo de colegio (variable stype) fijando $n_E = 100$ escuelas elementales, $n_M = 50$ escuelas medias y $n_H = 50$ escuelas superiores. De nuevo, nuestro interés recaerá sobre el número total de estudiantes matriculados y las medias de los índices api.

Veamos primero cómo obtener una muestra MAE con estas especificaciones. A continuación el código en R:

```
data(api)
attach(apipop)
table(stype)

## stype
##    E    H    M
## 4421  755 1018

set.seed(12345)
index = c(sample(which(stype=="E"),100),sample(which(stype=="H"),50),
sample(which(stype=="M"),50))
sample1 = apipop[index,]
```

Construyamos ahora una base de datos que contenga la muestra obtenida más el agregado de dos variables, una asociada a los pesos de muestreo (pw) y otra que especifique el tamaño del estrato que servirá para determinar el factor de corrección por finitud (fpc). En survey uno puede encontrar también una base de datos similar llamada apistrat. Nosotros llamaremos a nuestra base de datos sampleMAE.

```
aux = data.frame(pw = c(rep(44.21,100), rep(15.1,50), rep(20.36,50))
, fpc = c(rep(4421,100),rep(755,50), rep(1018,50)))
sampleMAE = cbind(sample1,aux)
```

Definamos ahora un objeto diseño apropiado que contenga la data y metada necesarias. Esto se hace con

```
disMAE = svydesign(id=~1,strata=~stype,fpc = ~fpc, data = sampleMAE)
```

Note que este comando tiene con respecto al MAS dos diferencias: una obvia `strata=~stype` que especifica la variable de estratificación y otra menos obvia dada por la introducción de la variable `fpc` del tamaño de la población en cada estrato. Si escribimos disMAE obtendremos:

```
disMAE

## Stratified Independent Sampling design
## svydesign(id = ~1, strata = ~stype, fpc = ~fpc, data = sampleMAE)
```

Analicemos ahora, como en el MAS, la estimación del número total de matriculados y la media del índice api para 1999 y el 2000:

```
svytotal(~enroll,disMAE,na.rm=T)

##          total     SE
## enroll 3831118 121207

svymean(~api99+api00,disMAE)

##       mean    SE
## api99  645 10.34
## api00  679  9.75
```

Como se aprecia, el MAE ha reducido (con relación al MAS) el error estándar de estimación. Esto es más evidente en la estimación del número total de estudiantes matriculados.

Mostremos ahora la estimación por dominios con la base de datos api y analicemos si bajo este MAE los colegios que cuentan con profesores con calificaciones de emergencia o no tienen un peor o mejor rendimiento api en el 2000. La variable `emer` recoge el porcentaje de profesores que tienen una calificación de emergencia en el colegio; es decir, de profesores que no han obtenido especialización en educación, pero que conocen del tema y que, por tanto, podrían brindar eventualmente las materias de su experticia. Note que cerca de un 80 % de

los colegios posee al menos un profesor con calificaciones de emergencia, lo cual indica la dificultad que tienen las escuelas para contratar profesores calificados.

```
table(as.numeric(apipop$emer>0))

##
##    0    1
## 1270 4922
```

Si deseáramos estimar manualmente bajo nuestro diseño la media del api en el 2000 y su error de estimación estimado, para el dominio de colegios que cuentan con algún profesor con calificaciones de emergencia, podríamos utilizar el siguiente código:

```
Nh = c(4421,755,1018)
nh = c(100,50,50)
disMAE = update(disMAE,cemer = as.integer(emer>0),apicemer = api00*cemer)
Nd_e = as.numeric(svytotal(~cemer,disMAE))
taud_e = as.numeric(svytotal(~apicemer,disMAE))
(Ybard_e = taud_e/Nd_e)

## [1] 652

zh = (sampleMAE$api00-Ybard_e)*(sampleMAE$emer>0)
sigma2hz = as.vector(by(zh,sampleMAE$stype,sd,na.rm=T))^2
(sed_e = sqrt(sum(Nh^2*(1-nh/Nh)*sigma2hz/nh)/Nd_e^2))

## [1] 10.6
```

Estos resultados pueden también obtenerse con el comando `subset` mediante

```
discemer = subset(disMAE,cemer==1)
svymean(~api00,discemer)

##       mean   SE
## api00  652 10.6
```

o de manera más general, para los dos dominios en estudio, con el comando `svyby` mediante

```
(mdom = svyby(~api00,~cemer,disMAE,svymean))

##   cemer api00   se
## 0     0   776 14.4
## 1     1   652 10.6
```

Como se observa, y puede verificarse a través de un intervalo de confianza, los colegios que contratan profesores con calificaciones de emergencia obtuvieron en promedio un menor índice de rendimiento api en el 2000 que aquellos que no contrataron a estos profesores.

3.6.2. MAE con la evaluación censal de estudiantes 2019

Consideremos un MAE para la población ECE 2019 de estudiantes del segundo año de secundaria de la DRE Amazonas. Algo primordial aquí es determinar el criterio de estratificación. Para ello, y tal como usualmente lo considera el Ministerio de Educación, usaremos como variables de estratificación a las definidas por el cruce entre las variables de Área (area) y Gestión (gestion2). Más específicamente, consideraremos 4 estratos: Urbana.Estatal, Rural.Estatal, Urbana.No estatal y Rural.No estatal. La construcción de la variable de estratificación y su distribución de frecuencias se muestra seguidamente.

```
load("ece19Am.RData")
Pop = ece19Am
Pop$Estrato=interaction(Pop$area,Pop$gestion2)
Pop = Pop[order(Pop$Estrato),]
table(Pop$Estrato)

##
##    Urbana.Estatal     Rural.Estatal Urbana.No estatal  Rural.No estatal
##              5324              2434                82                50
```

Supongamos ahora que deseamos estimar el rendimiento medio en Matemáticas para esta DRE con un error de no más de 5 puntos y una confianza del 95 %. Similarmente a como lo hicimos en el MAS, tomaremos para este fin una muestra piloto de tan solo 10 alumnos por estrato estimando con ello las desviaciones estándar iniciales por estrato. Esto también podría haberse hecho considerando las desviaciones estándar del rendimiento en Matemáticas en la ECE 2018 u otro estudio muestral previo para esta DRE. Véase por ejemplo el ejercicio 3.7.16.

```
set.seed(12345)
Nh = as.vector(table(Pop$Estrato))
sigmah=sd(Pop$M500_M[Pop$Estrato=="Urbana.Estatal"][sample(Nh[1],10)])
sigmah[2]=sd(Pop$M500_M[Pop$Estrato=="Rural.Estatal"][sample(Nh[2],10)])
sigmah[3]=sd(Pop$M500_M[Pop$Estrato=="Urbana.No estatal"][sample(Nh[3],10)])
sigmah[4]=sd(Pop$M500_M[Pop$Estrato=="Rural.No estatal"][sample(Nh[4],10)])
```

Utilizaremos aquí la siguiente afijación de Neyman a_h para el tamaño de muestra por estrato

```
ah = Nh*sigmah/sum(Nh*sigmah)
```

Los tamaños de muestra vendrán dados entonces por

```
d = dim(Pop)[1]*5/qnorm(0.975)
n = sum(((Nh*sigmah)^2)/ah)/(d^2 + sum(Nh*sigmah^2))
(n = ceiling(n))

## [1] 1020

(nh = round(ah*n))

## [1] 646 353  15   5
```

Tomemos ahora la muestra donde, a diferencia del ejemplo anterior, utilizaremos el comando `strata` del paquete sampling. Este nos permitirá obtener la muestra de una manera mucho más directa. Tal comando, cabe comentar, requiere de una previa ordenación (como lo hicimos) de la base de datos según la variable de estratificación.

```
library(sampling)
set.seed(12345)
m=strata(Pop,c("Estrato"),size=nh,method="srswor")
me19Am = getdata(Pop,m)
table(is.na(me19Am$M500_M))

##
## FALSE  TRUE
##  1014     5
```

Puesto que nuestra intención es analizar los rendimientos en Matemáticas y tenemos aquí casos perdidos, eliminemos primero estos de la muestra:

```
me19Am = me19Am[is.na(me19Am$M500_M)==0,]
nh = as.vector(table(me19Am$Estrato))
nh

## [1] 645 349  15   5

me19Am = cbind(me19Am,fpc = rep(Nh,nh))
```

El objeto diseño será entonces:

```
dis19MAE = svydesign(id=~1,strata=~Estrato,fpc=~fpc,data=me19Am)
```

Este nos dará las siguientes estimaciones para los rendimientos medios y la proporción de logros alcanzados en Matemáticas por los estudiantes del segundo año de secundaria en Amazonas:

```
(meanEAm = svymean(~M500_M,dis19MAE, deff=T))

##          mean     SE DEff
## M500_M 527.79   2.88 0.92

(mpM = svymean(~grupo_M,dis19MAE,na.rm=T))

##                             mean   SE
## grupo_MPrevio al inicio 0.4597 0.01
## grupo_MEn inicio        0.2854 0.01
## grupo_MEn proceso       0.1572 0.01
## grupo_MSatisfactorio    0.0978 0.01
```

Note que para el rendimiento medio en Matemáticas pedimos una estimación del efecto del diseño, el cual, como se aprecia, demuestra una ligera mayor eficiencia del MAE en comparación con el MASs.

3.6.3. MAE para la población penitenciaria 2016

Supongamos que en lugar del censo penitenciario 2016 se nos hubiese encargado en tal fecha diseñar un muestreo aleatorio estratificado para la población penitenciaria del país. La pregunta inicial es entonces cómo considerar los estratos. Recordemos que un MAE es óptimo mientras más pueda separar a la población en estratos relativamente homogéneos. Claramente, una variable con tal propiedad para nuestra población penitenciaria es el sexo;

otra podría ser el nivel de peligrosidad de los internos, el cual lamentablemente desconocemos. Una posible tercera variable discriminatoria podría ser el nivel de hacinamiento de las cárceles, información que si bien no está consignada en el censo, es posible obtener de conocerse la capacidad de los establecimientos penitenciarios (EP). Como criterio, consideraremos que un EP se encuentra en condición de hacinamiento si este alberga al doble o más de internos de su capacidad. Cabe precisar que las variables de capacidad y sexo para los EP se obtuvieron de una fuente externa al censo (Informe Estadístico Penitenciario Noviembre 2016. INPE). El siguiente código define los posibles estratos que resultarán del cruce de las variables de condición de género y hacinamiento.

```
load('cp16.RData')
ncap = c(888,65,50,150,72,1518,160,1143,1370,50,384,350,920,572,1152,1464,
768,823,644,1620,2200,288,450,548,42,3204,1142,667,67,78,222,40,214,644,42,
60,120,680,105,85,50,48,64,1074,96,788,90,248,800,62,80,590,288,60,286,600,
78,654,544,636,180,44,778,420,1,8)
sex = c(3,1,3,1,2,1,2,3,1,2,3,3,3,1,1,3,1,1,3,1,3,2,2,2,1,1,1,1,2,1,1,2,1,3,
1,1,3,1,2,2,1,1,3,3,3,3,3,3,1,2,3,1,3,3,3,1,2,3,3,1,1,2,1,3,1,1)
freq = as.vector(table(cp16$EST_PENIT))
phacib = freq/ncap
hacib = as.numeric(phacib>=2)
table(hacib,sex)

##      sex
## hacib  1  2  3
##     0 12  9  7
##     1 17  4 17
```

La tabla final muestra la distribución de EP por condición de hacinamiento y sexo; sin embargo, en lugar de trabajar con estos potenciales 6 estratos, creemos que sería más conveniente considerar tan solo 4, ya que los EP mixtos (EP que albergan tanto a hombres como mujeres) podríamos subdividirlos en dos EP: una para hombres y otro para mujeres. De esta manera incrementaríamos la cantidad de EP de 24 a 90. Todas estas correcciones y actualizaciones de la base de datos se muestran a través del siguiente código:

```
# Recodificación de los establecimientos penitenciarios
aux = levels(cp16$EST_PENIT)
EP=factor(cp16$EST_PENIT,levels=c(aux[c(2,4:7,9,10,14,15,17,18,20,22:33,35,36,38:42,49,50,
52,56,57,60:63,65,66)],"Cajamarca_h","Cajamarca_m","Jaen_h","Jaen_m","Chiclayo_h",
"Chiclayo_m","Tumbes_h","Tumbes_m","Huaraz_h","Huaraz_m","Chimbote_h","Chimbote_m",
"Ica_h","Ica_m","Huacho_h","Huacho_m","Ancon2_h","Ancon2_m","Ayacucho_h","Ayacucho_m",
"Chanchamayo_h","Chanchamayo_m","Oroya_h","Oroya_m","Huanuco_h","Huanuco_m",
```

```
"Cerro Pasco_h","Cerro Pasco_m","Pucallpa_h","Pucallpa_m","Abancay_h","Abancay_m",
"Andahuaylas_h","Andahuaylas_m","Quillabamba_h","Quillabamba_m","Chachapoyas_h",
"Chachapoyas_m","Bagua Grande_h","Bagua Grande_m","Yurimaguas_h","Yurimaguas_m",
"Juanjui_h","Juanjui_m","Moyobamba_h","Moyobamba_m","Juliaca_h","Juliaca_m"))
EP[cp16$EST_PENIT=="Cajamarca" & cp16$GENERO=="Hombre"] <- "Cajamarca_h"
EP[cp16$EST_PENIT=="Cajamarca" & cp16$GENERO=="Mujer"] <- "Cajamarca_m"
EP[cp16$EST_PENIT=="Jaen" & cp16$GENERO=="Hombre"] <- "Jaen_h"
EP[cp16$EST_PENIT=="Jaen" & cp16$GENERO=="Mujer"] <- "Jaen_m"
EP[cp16$EST_PENIT=="Chiclayo" & cp16$GENERO=="Hombre"] <- "Chiclayo_h"
EP[cp16$EST_PENIT=="Chiclayo" & cp16$GENERO=="Mujer"] <- "Chiclayo_m"
EP[cp16$EST_PENIT=="Tumbes" & cp16$GENERO=="Hombre"] <- "Tumbes_h"
EP[cp16$EST_PENIT=="Tumbes" & cp16$GENERO=="Mujer"] <- "Tumbes_m"
EP[cp16$EST_PENIT=="Huaraz" & cp16$GENERO=="Hombre"] <- "Huaraz_h"
EP[cp16$EST_PENIT=="Huaraz" & cp16$GENERO=="Mujer"] <- "Huaraz_m"
EP[cp16$EST_PENIT=="Chimbote" & cp16$GENERO=="Hombre"] <- "Chimbote_h"
EP[cp16$EST_PENIT=="Chimbote" & cp16$GENERO=="Mujer"] <- "Chimbote_m"
EP[cp16$EST_PENIT=="Ica" & cp16$GENERO=="Hombre"] <- "Ica_h"
EP[cp16$EST_PENIT=="Ica" & cp16$GENERO=="Mujer"] <- "Ica_m"
EP[cp16$EST_PENIT=="Huacho" & cp16$GENERO=="Hombre"] <- "Huacho_h"
EP[cp16$EST_PENIT=="Huacho" & cp16$GENERO=="Mujer"] <- "Huacho_m"
EP[cp16$EST_PENIT=="Modelo Ancon II - S.M.V.C." & cp16$GENERO=="Hombre"] <- "Ancon2_h"
EP[cp16$EST_PENIT=="Modelo Ancon II - S.M.V.C." & cp16$GENERO=="Mujer"] <- "Ancon2_m"
EP[cp16$EST_PENIT=="Ayacucho" & cp16$GENERO=="Hombre"] <- "Ayacucho_h"
EP[cp16$EST_PENIT=="Ayacucho" & cp16$GENERO=="Mujer"] <- "Ayacucho_m"
EP[cp16$EST_PENIT=="Chanchamayo" & cp16$GENERO=="Hombre"] <- "Chanchamayo_h"
EP[cp16$EST_PENIT=="Chanchamayo" & cp16$GENERO=="Mujer"] <- "Chanchamayo_m"
EP[cp16$EST_PENIT=="Oroya" & cp16$GENERO=="Hombre"] <- "Oroya_h"
EP[cp16$EST_PENIT=="Oroya" & cp16$GENERO=="Mujer"] <- "Oroya_m"
EP[cp16$EST_PENIT=="Huanuco" & cp16$GENERO=="Hombre"] <- "Huanuco_h"
EP[cp16$EST_PENIT=="Huanuco" & cp16$GENERO=="Mujer"] <- "Huanuco_m"
EP[cp16$EST_PENIT=="Cerro Pasco" & cp16$GENERO=="Hombre"] <- "Cerro Pasco_h"
EP[cp16$EST_PENIT=="Cerro Pasco" & cp16$GENERO=="Mujer"] <- "Cerro Pasco_m"
EP[cp16$EST_PENIT=="Pucallpa" & cp16$GENERO=="Hombre"] <- "Pucallpa_h"
EP[cp16$EST_PENIT=="Pucallpa" & cp16$GENERO=="Mujer"] <- "Pucallpa_m"
EP[cp16$EST_PENIT=="Abancay" & cp16$GENERO=="Hombre"] <- "Abancay_h"
EP[cp16$EST_PENIT=="Abancay" & cp16$GENERO=="Mujer"] <- "Abancay_m"
EP[cp16$EST_PENIT=="Andahuaylas" & cp16$GENERO=="Hombre"] <- "Andahuaylas_h"
EP[cp16$EST_PENIT=="Andahuaylas" & cp16$GENERO=="Mujer"] <- "Andahuaylas_m"
EP[cp16$EST_PENIT=="Quillabamba" & cp16$GENERO=="Hombre"] <- "Quillabamba_h"
EP[cp16$EST_PENIT=="Quillabamba" & cp16$GENERO=="Mujer"] <- "Quillabamba_m"
EP[cp16$EST_PENIT=="Chachapoyas" & cp16$GENERO=="Hombre"] <- "Chachapoyas_h"
EP[cp16$EST_PENIT=="Chachapoyas" & cp16$GENERO=="Mujer"] <- "Chachapoyas_m"
EP[cp16$EST_PENIT=="Bagua Grande" & cp16$GENERO=="Hombre"] <- "Bagua Grande_h"
EP[cp16$EST_PENIT=="Bagua Grande" & cp16$GENERO=="Mujer"] <- "Bagua Grande_m"
EP[cp16$EST_PENIT=="Yurimaguas" & cp16$GENERO=="Hombre"] <- "Yurimaguas_h"
EP[cp16$EST_PENIT=="Yurimaguas" & cp16$GENERO=="Mujer"] <- "Yurimaguas_m"
```

```
EP[cp16$EST_PENIT=="Juanjui" & cp16$GENERO=="Hombre"] <- "Juanjui_h"
EP[cp16$EST_PENIT=="Juanjui" & cp16$GENERO=="Mujer"] <- "Juanjui_m"
EP[cp16$EST_PENIT=="Moyobamba" & cp16$GENERO=="Hombre"] <- "Moyobamba_h"
EP[cp16$EST_PENIT=="Moyobamba" & cp16$GENERO=="Mujer"] <- "Moyobamba_m"
EP[cp16$EST_PENIT=="Juliaca" & cp16$GENERO=="Hombre"] <- "Juliaca_h"
EP[cp16$EST_PENIT=="Juliaca" & cp16$GENERO=="Mujer"] <- "Juliaca_m"
cp16 = cbind(cp16,EP)
# Agregando las nuevas variables capacidad, CG, hacinamiento y estrato
cp16  = cp16[order(cp16$EST_PENIT),]
Capacidad = rep(ncap,freq)
CG = rep(sex,freq)
Hac = rep(hacib,freq)
cp16f = cbind(cp16,Capacidad,CG,Hac)
cp16f$Estrato=1
cp16f$Estrato[cp16f$Hac==0 & cp16f$GENERO=="Mujer"] = 2
cp16f$Estrato[cp16f$Hac==1 & cp16f$GENERO=="Hombre"] = 3
cp16f$Estrato[cp16f$Hac==1 & cp16f$GENERO=="Mujer"] = 4
cp16f$Estrato=factor(cp16f$Estrato,labels=c("No hacinados hombres","No hacinados mujeres",
"Hacinados hombres","Hacinados mujeres"))
cp16f  = cp16f[order(cp16f$Estrato),]
save(cp16f,file='cp16f.RData')
```

Una mirada parcial a la base de datos final nos revela lo siguiente:

```
cp16f[1:8,c(7:13,190:194)]

##    EST_PENIT PABELLON GENERO     E_CIVIL   RELIGION EDAD    NACIONALIDAD
## 3  Cajamarca       NA Hombre   Casado(a)    Ninguna   25 PERUANO
## 4  Cajamarca       NA Hombre    Viudo(a)       Otra   26 PERUANO
## 5  Cajamarca        3 Hombre   Casado(a) Evangélica   49 PERUANO
## 6  Cajamarca        7 Hombre Conviviente    Ninguna   40 PERUANO
## 7  Cajamarca        1 Hombre   Casado(a)   Católica   25 PERUANO
## 8  Cajamarca        2 Hombre   Casado(a)   Católica   45 PERUANO
## 10 Cajamarca       NA Hombre Conviviente   Católica   40 PERUANO
## 11 Cajamarca       NA Hombre   Casado(a) Evangélica   40 PERUANO
##             EP Capacidad CG Hac              Estrato
## 3  Cajamarca_h       888  3   0 No hacinados hombres
## 4  Cajamarca_h       888  3   0 No hacinados hombres
## 5  Cajamarca_h       888  3   0 No hacinados hombres
## 6  Cajamarca_h       888  3   0 No hacinados hombres
## 7  Cajamarca_h       888  3   0 No hacinados hombres
## 8  Cajamarca_h       888  3   0 No hacinados hombres
## 10 Cajamarca_h       888  3   0 No hacinados hombres
```

```
## 11 Cajamarca_h       888  3   0 No hacinados hombres
```

Definamos ahora el diseño MAE y tomemos la muestra. Para ello usaremos el mismo tamaño de muestra del MASs con una asignación proporcional, lo que nos da para cada estrato tamaños de muestra de, respectivamente, 152, 37, 838 y 27 internos.

```
set.seed(12345)
Nh = as.numeric(table(cp16f$Estrato))
m = strata(cp16f,c("Estrato"),size=c(152,37,838,27),method="srswor")
sampleMAE = getdata(cp16f,m)
fpc=c(rep(Nh[1],152),rep(Nh[2],37), rep(Nh[3],838),rep(Nh[4],27))
sampleMAE=cbind(sampleMAE,fpc)
disenhoMAE = svydesign(id=~1,strata=~Estrato,fpc = ~fpc, data = sampleMAE)
disenhoMAE

## Stratified Independent Sampling design
## svydesign(id = ~1, strata = ~Estrato, fpc = ~fpc, data = sampleMAE)
```

Estimemos, finalmente, como en el capítulo 2, la edad promedio de los internos, la proporción de internos sentenciados y la proporción de aquellos que cuentan con un abogado.

```
svymean(~EDAD, disenhoMAE,na.rm=T)

##      mean   SE
## EDAD 36.1 0.35

svymean(~SITUACION_JURIDICA,disenhoMAE,na.rm=T)

##                                  mean   SE
## SITUACION_JURIDICAProcesado   0.206 0.01
## SITUACION_JURIDICASentenciado 0.794 0.01

svymean(~ABOGADO,disenhoMAE,na.rm=T)

##            mean   SE
## ABOGADOSí 0.533 0.02
## ABOGADONo 0.467 0.02
```

3.7. Ejercicios

1. Se desea estimar la media poblacional de una variable y mediante un MAE. Muestre que la varianza estimada de su estimador insesgado bajo la asignación de Neyman es siempre menor o igual que la de este estimador mediante la asignación proporcional y muestre que

$$\hat{V}_{Prop}(\bar{Y}) - \hat{V}_{Neyman}(\bar{Y}) = \frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}(S_h - \sum_{j=1}^{H}\frac{N_j}{N}S_j)^2.$$

Explicíte esta diferencia para $H = 2$ e indique qué debería ocurrir para que esta diferencia sea cada vez más grande.

2. Considere un MAE de tamaño n de solo 2 estratos en el que es de interés estimar la diferencia de medias de una variable y entre estos estratos.
a) Proponga un estimador insesgado para esta diferencia μ_D y un estimador para su error estándar de estimación.
b) ¿Qué asignación de tamaños de muestra por estrato haría que el error estándar de estimación de μ_D sea mínimo?
c) En una encuesta por MAE de 300 trabajadores de una universidad, con estratos definidos por quienes tienen educación superior y no, es de interés comparar el tiempo medio en horas por día que estos grupos de trabajadores permanecen en la universidad. Un estudio piloto previo encontró que quienes tenían estudios superiores presentaron una media y desviación estándar de 8.25 y 3.46 horas por día, respectivamente; mientras que los que no tenían estudios superiores tuvieron un media y desviación estándar de 7.45 y 4.122, respectivamente. Halle la asignación óptima de los 300 trabajadores, que se debería de tener por estrato, de tal manera que se obtenga un intervalo de confianza de mínima longitud esperada para la diferencia de medias en discusión.

3. Si se realiza un MAE para una población con 3 estratos de 50, 80 y 70 unidades, ¿cuántas muestras distintas de tamaño 40 podrían obtenerse bajo una asignación proporcional?

4. Dado los resultados de un MAE, muestre que un estimador insesgado de la varianza de la media bajo un MASs

$$V_{MASs}(\bar{Y}) = (1 - \frac{n}{N})\frac{\sigma^2_{N-1}}{n}$$

viene dado por

$$\hat{V}_{MASs}(\bar{Y}) = \frac{(N-n)}{n(N-1)}(\frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{N_h}y_{hi}^2\delta_{hi} - \bar{Y}^2 + \hat{V}(\bar{Y})),$$

donde $\hat{V}(\bar{Y})$ viene dado por (3.1).

5. Muestre que el estimador insesgado de la varianza de la media de una variable y bajo una asignación proporcional en el ejercicio 4 toma la forma

$$\hat{V}_{MASs}(\bar{Y}) = \frac{(N-n)}{n(N-1)}(\frac{n-1}{n}S^2 + \hat{V}(\bar{Y})),$$

donde S^2 denota la varianza de todos los datos en la muestra sin tomar en cuenta la estratificación. Obtenga esta estimación para los datos de la ECE 2019 de Amazonas si se toma en ella un MAE con asignación proporcional de tamaño 1000 que busca estimar el rendimiento medio en Matemáticas de los alumnos del segundo año de secundaria.

6. Divida los rectángulos del ejercicio 20 del capítulo 2 en 2 estratos según estos tengan o no menos de 30 unidades. Tome luego una MAE de tamaño 20 con algún tipo de afijación para estimar el área total de todos los rectángulos y su intervalo de confianza al 98 %. Compare sus resultados con los del ejercicio 20.

7. Una biblioteca municipal desea estimar el porcentaje de libros infantiles que posee. La biblioteca tiene 4 salas (Norte, Sur, Este y Oeste) con 2800, 2940, 4050 y 7900 libros, respectivamente. Suponga que para este objetivo se tomó un MAE con un tamaño de muestra correspondiente al 10 % de los libros de la biblioteca y una asignación proporcional.
a) Si en la sala Sur se ubicaron 30 libros infantiles, ¿cuál es la estimación de la proporción de libros infantiles en dicha sala?
b) Para un nivel de confianza del 98 %, ¿cuál es el error máximo de estimación que se está cometiendo en la estimación anterior?
c) A un nivel de confianza del 95 %, ¿qué error máximo de estimación reportaría usted al estimar la proporción de libros infantiles en toda la biblioteca?

8. En ocasiones, un MAE puede no estar adecuadamente equilibrado en alguna variable no considerada como criterio de estratificación o podríamos no saber cuándo una unidad proviene de un estrato particular sino hasta después de observar la muestra. Por posestratificación entenderemos al proceso mediante el cual uno extrae un MAS de la población y estratifica esta luego de ser observada. En consecuencia, los tamaños de muestra en cada (pos)estrato resultan aleatorios. Si para la media poblacional μ de una variable y consideramos al estimador $\bar{Y} = \sum_{h=1}^{H} \frac{N_h}{N}\bar{Y}_h$, donde $\bar{Y}_h$ es la media muestral de y en el (pos)estrato h, N_h el número de unidades en este (post)estrato y asumimos un muestreo sin reemplazamiento,
a) Muestre que tanto los $\bar{Y}_h$ como $\bar{Y}$ son estimadores insesgados de, respectivamente, μ_h y μ, siendo μ_h la media poblacional de y en el (pos)estrato h y μ la media poblacional global.
b) Muestre que $V(\bar{Y}) = \sum_{h=1}^{H}(\frac{N_h}{N})^2\sigma_h^2(E(\frac{1}{X_h})-\frac{1}{N_h})$, donde σ_h^2 y X_h denotan, respectivamente, la varianza y el tamaño de muestra en el (pos)estrato h.
c) Proponga algún estimador insesgado de N_h, en caso de que este sea desconocido.
d) Use una expansión de Taylor de segundo orden para el valor esperado en b) y muestre que una aproximación de la varianza de $\bar{Y}_h$ y de la varianza de $\bar{Y}$ vienen dadas, respectivamente,

por

$$V(\bar{Y}_h) = (1 + \frac{(N - N_h)N}{nN_h(N-1)})\frac{(N-n)\sigma_h^2}{nN_h} \quad \text{y}$$

$$V(\bar{Y}) = \frac{N-n}{nN}\sum_{h=1}^{H}(\frac{N_h}{N})\sigma_h^2 + \frac{1}{n^2}(\frac{N-n}{N-1})\sum_{h=1}^{H}(\frac{N-N_h}{N})\sigma_h^2.$$

e) Dos estimadores de $V(\bar{Y})$ son los estimadores incondicionales y condicionales. El primero se obtiene simplemente al estimar insesgadamente en él las varianzas de los (pos)estratos o los N_h, de ser necesarios, y el segundo de igual manera pero omitiendo el segundo término a derecha en esta varianza. Obtenga estas estimaciones si al tomar el MASs en 2.4.3 quisieramos estimar la media del índice api del 2000, pero haciendo ahora una posestratificación según sea el colegio elegible o no para reconocimientos (*awards*). Obtenga, finalmente, una estimación de $V(\bar{Y})$ mediante el paquete survey, utilizando para ello el comando postStratify.

9. La DRE de Lima Metropolitana le ha pedido a usted realizar un MAE para la población ECE 2019 con los siguientes 3 estratos: alumnos de colegios urbanos estatales, alumnos de colegios urbanos no estatales y alumnos de colegios del área rural. Su presupuesto le alcanza para evaluar 3000 alumnos y tiene como objetivo estimar el rendimiento medio en Matemáticas de esta DRE. Si utiliza una asignación proporcional,
a) ¿Qué problemas prácticos piensa que podría encontrar al momento de realizar el muestreo?
b) ¿Cuál sería el error de estimación que reportaría en este estudio bajo un nivel de confianza del 95 % ?
c) Si le piden, como parte del estudio, reportar también el rendimiento medio en Matemáticas, según el nivel socieconómico, donde estos niveles son 3 y definidos por quienes tienen un ISE menor a 0.4, entre 0.4 y 1 y mayores a 1, ¿cuáles serían las estimaciones pedidas y sus errores estándar de estimación estimados?

10. Un instituto cuenta con las especialidades de Contabilidad, Diseño Industrial, Arquitectura de interiores y Administración de Negocios y desea estimar la proporción de sus egresados que estarían dispuestos a seguir una nueva diplomatura que el instituto piensa abrir. Se sabe que el último año egresaron de estas especialidades, respectivamente, 20, 200, 80 y 230 alumnos. Si le informan que una encuesta de 50 egresados ya ha sido tomada mediante un MASs,
a) ¿Qué tan probable es que la encuesta haya omitido a alguna especialidad?
b) Si en la muestra se obtuvo, respectivamente, 3, 20, 12 y 15 alumnos de cada una de las especialidades anteriores, donde 2, 4, 7 y 8 de ellos manifestaron que seguirían el diploma, use esta información como muesta piloto para encontrar en un estudio futuro el tamaño de muestra que se requeriría en un MAE para estimar la proporción de aceptación del diploma con un margen de error de 0.03 y un nivel de confianza del 95 %. Proponga el tipo de asignación.

11. Considere una población de $N = 20$ domicilios, donde es conocida la variable y = renta familiar mensual en miles de soles y la variable estrato socioeconómico al cual pertenecen (con niveles A = alto, M = medio y B = bajo). Los valores de estas variables se resumen en la siguiente tabla:

Unidad	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Renta	13	17	6	5	9	12	19	6	14	12	8	5	11	20	6	18	10	9	12	8
Estrato	M	A	B	B	B	M	A	B	M	M	B	B	M	A	B	A	M	B	B	B

A fin de estimar la renta familiar media, se tienen las alternativas de efectuar un MAE con afijación proporcional, un MASs o un MASc; todos de tamaño 10.

a) Determine las varianzas de estos diseños e indique cuál es más eficiente y por qué.

b) Usando los números aleatorios 0.91, 0.02, 0.7, 0.35, 0.1, 0.96, 0.51, 0.46, 0.23, 0.87, tome las muestras requeridas para estos diseños y estime la renta familiar media bajo cada uno.

12. Una empresa desea estimar, con un error no mayor a las 250 horas y un nivel de confianza del 95 %, el número total de horas de trabajo perdidas en un mes debido a accidentes entre sus empleados (tomado de Mendenhall et al. (2007)). Como los obreros, técnicos y administradores tienen diferentes tasas de siniestralidad, el investigador decide utilizar un muestreo aleatorio estratificado, de modo que cada grupo forma un estrato independiente. Los datos de años anteriores sugieren las varianzas que se muestran abajo por el número de horas de trabajo perdidas por empleado en los 3 grupos. Se muestran también los tamaños actualizados de los estratos.

	Obreros	Técnicos	Administradores
Varianza	36	25	9
Tamaño	132	92	27

Usando la afijación de Neyman, determine los tamaños de muestra adecuados. ¿Cambian estos tamaños si la asignación es proporcional?

13. En el MAE hemos seguido siempre la estrategia de obtener los tamaños de muestra según las especificaciones del máximo error de estimación tolerable para estimar un parámetro poblacional a un nivel de confianza dado. En ciertas situaciones, sin embargo, el investigador podría estar interesado en tratar de estimar el parámetro de interés para cada estrato con un máximo error de estimación prefijado en él a un nivel de confianza dado. La pregunta entonces es ¿cuál es el máximo error de estimación que se estaría cometiendo al estimar con este procedimiento el parámetro en toda la población para el nivel de confianza dado? Resuelva este problema para el caso del ejercicio 10; asuma que en este se desee estimar el número total de horas de trabajo perdidas al interior de cada estrato con un error no mayor a las 100 horas y una confianza del 95 %.

14. Suponga que en el MAE de la subsección 3.6.2 para el ECE 2019 de Amazonas le piden que reporte las estimaciones del rendimiento medio en Matemáticas por sexo.
a) Dé estas estimaciones y sus errores estándar de estimación estimados.
b) ¿Cómo haría para comparar el rendimiento medio de las estudiantes mujeres que pertenecen a colegios estatales y no estatales? ¿Se podría concluir, con una confianza del 95 %, que hay diferencias entre estos rendimientos medios?
c) Responda b) para el caso de los estudiantes hombres.

15. Burnard (1992) envió un cuestionario a una muestra estratificada de tutores y estudiantes en Gales para estudiar lo que ellos entendían por el término experiential learning. Los tamaños de población y muestra de los cuatro estratos se muestran a continuación:

Estrato	Tamaño de la población	Tamaño de muestra
Tutores generales de enfermería (GT)	150	109
Tutores de enfermería psiquiátrica (PT)	34	26
Estudiantes generales de enfermería (GS)	2680	222
Estudiantes de enfermería psiquiátrica (PS)	570	40

A los entrevistados se les preguntó cuáles de las siguientes técnicas podrían identificarlas como métodos de *experiential learning*. El número de entrevistados de cada grupo que identificó el método como de *experiential learning* se muestra a continuación:

Método	GS	PS	PT	GT
Role play	213	38	26	104
Problemas de solución de actividades	182	33	22	95
Simulaciones	95	20	22	64
Empatía en la construcción de ejercicios	89	25	20	54
Ejercicios gestálquicos	24	4	5	12

Estime el porcentaje total de estudiantes de enfermería y tutores que identifican cada una de las técnicas mencionadas como de *experiential learning*. Indique también en cada caso el error estándar de estimación estimado para cada una de sus estimaciones.

16. Considere un MAE con asignación óptima sobre la ECE 2019 del segundo año de secundaria de la DRE Cusco; use la misma estratificación que en el estudio de Amazonas y estime su rendimiento medio en Matemáticas. Se pide para ello un error de estimación no mayor a los 5 puntos con una confianza del 95 %.
a) ¿Qué tamaño de muestra debería considerar para este dominio? Para responder utilice la ECE 2018.
b) Tome la muestra requerida y obtenga la estimación pedida calculando así como la estimación del efecto de diseño.
c) Compare, mediante un intervalo de confianza al 95 %, los rendimientos medios de Matemáticas entre Cusco y Amazonas.

17. Considere la base de datos poblacional Province91 del ejercicio 15 del capítulo 2 y la variable Stratum allí definida que identifica si la municipalidad de la provincia en estudio es urbana o rural. Usando esta última como variable de estratificación y la variable número de personas desempleadas como variable de investigación, tome un MAE de 8 municipalidades y responda a lo siguiente:
a) Halle los tamaños de muestra por estrato usando una asignación proporcional.
b) Obtenga para el diseño anterior los efectos de diseño en la estimación del total de personas desempleadas de la provincia.
c) Tomando la muestra requerida, estime tanto el total de personas desempleadas en la provincia como el efecto de diseño en esta estimación.
d) Si se estimara, bajo este diseño, la proporción de municipalidades que tienen una población económicamente activa superior a las 10 000 personas, ¿qué error de estimación estimaría para esta proporción?

18. Considere la base de datos apipop y suponga que está interesado en estimar el número total de alumnos matriculados en esta población con un MAE, donde el criterio de estratificación sea nuevamente el tipo de colegio. Se desea estimar este número con un error de estimación no mayor a los 70 000 alumnos y un nivel de confianza del 95 %. Para ello,
a) Tome un MAE piloto de solo 30 escuelas, usando por simplicidad una asignación proporcional, e indique en cúanto estimaría las desviaciones estándar del número de matriculados por tipo de colegio.
b) Halle los tamaños de muestra requeridos con una asignación óptima y costos de muestreo iguales utilizando las estimaciones necesarias de la muestra piloto tomada en a).
c) Realice el MAE y reporte el IC al 95 % para el número de matriculados en esta población.
d) Si en la muestra anterior era también de interés estimar la proporción de escuelas en esta población que recibieron un premio (*awards*), estime tal proporción y reporte su error de estimación estimado.

19. Un hospital público está interesado en construir en sus instalaciones una clínica privada y por ello desea realizar una encuesta por muestreo para estimar, entre otras cosas, la proporción de familias de la ciudad que se atenderían en esta clínica. El diseño sugerido será estratificado y se tomarán como variables de estratificación a una que indique si la familia utiliza o ha utilizado el hospital o no lo ha hecho y a otra que indica si la familia proviene del distrito donde se ubica el hospital o no. Los cuatro estratos formados, que denotaremos como 1, 2, 3 y 4, serán entonces los de las familias usuarias del distrito, las usuarias que no son del distrito, por las no usuarias del distrito y por las no usuarias que no son del distrito. En un estudio piloto se encontró que, aproximadamente, el 85 % de los usuarios y el 45 % de los nos usuarios se atenderían en la clínica; sin embargo el estudio piloto no registró el distrito de residencia del entrevistado. Algo que tomar en cuenta en la encuesta será que el costo de obtener una observación para un usuario será de 5 soles; mientras que el costo para

un no usuario será de 9 soles, pues los no usuarios son más difíciles de ubicar. Además, se sabe que el número de familias en estos estratos, según el último censo, son de $N_1 = 123$, $N_2 = 65$, $N_3 = 155$ y $N_4 = 570$.
a) Encuentre qué proporción óptima de la muestra total debería corresponder a cada estrato. ¿Qué criterio está utilizando para obtener estas asignaciones?
b) Encuentre el tamaño de muestra total si se desea estimar la proporción buscada con un error de estimación no mayor a 0.05 y un nivel de confianza del 95 %.
c) Suponga que tiempo después de realizado el estudio se encontró para cada estrato las siguientes estimaciones de la proporción de familias que usarían la clínica

$$\hat{p}_1 = 0.85, \quad \hat{p}_2 = 0.92, \quad \hat{p}_3 = 0.55, \quad \hat{p}_4 = 0.43.$$

Estime la proporción poblacional p buscada y su error estándar de estimación.
d) Si el presupuesto total para el muestreo se hubiese fijado en \$400, ¿cuáles serían ahora los tamaños de muestra por estrato que minimicen el error de estimación?
e) Suponga que le pidiesen ahora que en cada estrato el error de estimación en la proporción de familias que se atenderían en la clínica no fuese mayor a 0.05, con una confianza del 95 %. ¿Cuál sería el tamaño de muestra total que se requeriría en este muestreo?

20. En esta actividad se le pide que realice un MAE para la base de datos de libros en línea de la compañía Amazon (EE.UU.). El trabajo se restringirá a solo la población de libros de Estadística (Statistics) que no estén fuera de *stock* y que sean nuevos. El criterio de estratificación se basará en el formato o tipo de empastado que tienen los libros (*paperback, hardcover, loose leaf* y otros). Usando una asignación proporcional y una muestra de tamaño 70, estime, junto con sus errores estándar de estimación, el precio medio, la puntuación media y la proporción de libros del 2017 para cada estrato y para la población en general. Repita este ejercicio para el año actual.

21. Arias-Schreiber et al. (2019) realizaron un análisis de costo-beneficio (ACB) a los proyectos de ley presentados solo por congresistas en el período legislativo 2012-2013 del Congreso de la República del Perú (980 proyectos de ley entre el 27/07/2012 al 15/06/2013). Ellos construyeron un indicador de calidad del análisis ACB sobre la base de 18 variables que medían distintas características de los proyectos de ley, características tales como la identificación de los beneficiarios y perjudicados por el proyecto, la necesidad de presentar el proyecto, el uso de información para sustentar el proyecto, la evaluación de los costos para el Estado de aprobarse el proyecto, etc. Información sobre estas variables puede obtenerse en la siguiente página web del Congreso de la República:

`http://www2.congreso.gob.pe/Sicr/TraDocEstProc/CLProLey2011.nsf/`

Para este análisis se optó por tomar un MAE con tres estratos que reflejaban la participación de la comisión adscrita al proyecto en el presupuesto del Estado. Concretamente, cada uno de los estratos que se formaron fueron los siguientes:

- (1) Estrato I (hasta el 1 % del presupuesto). Comprendió a las comisiones de Comercio Exterior y Turismo, Constitución y Reglamento, Cultura y Patrimonio Cultural, Energía y Minas, Fiscalización y Contraloría, Inclusión Social y Personas con Discapacidad, Mujer y Familia, Producción, Micro y Pequeña Empresa y Cooperativas, Pueblos Andinos, Amazónicos y Afroperuanos, Ambiente y Ecología, Relaciones Exteriores, y Trabajo y Seguridad Social. Se registraron 318 proyectos de ley en este estrato.
- (2) Estrato II (por encima del 1 % y hasta el 6 % del presupuesto). Comprendió a las comisiones Agraria, de Ciencia, Innovación y Tecnología de Defensa del Consumidor y Organismos Reguladores de los Servicios Públicos, de Justicia y Derechos Humanos, de Salud y Población y de Vivienda y Construcción. Se registraron 319 proyectos de ley en este estrato.
- (3) Estrato III (por encima del 6 % y hasta el 22 % del presupuesto) Comprendió a las comisiones de Defensa Nacional, Orden Interno, Desarrollo Alternativo y Lucha contra las Drogas, de Descentralización, Regionalización, Gobiernos Locales y Modernización de la Gestión del Estado, de Economía, Banca y Finanzas e Inteligencia Financiera, de Educación, Juventud y Deporte, de Transportes y Comunicaciones, de Inteligencia y de Presupuesto y Cuenta General de la República. Se registraron 343 proyectos de ley en este estrato.

Si se desea estimar el porcentaje de proyectos de ley que tuvieron un análisis ACP aceptable con un nivel de confianza del 95 % y un margen de error del 10 %,
a) ¿Cuál sería el tamaño de muestra adecuado? Tome en cuenta que, según estudios previos del ACB en el país y en países de la región, este porcentaje nunca supero el 20 %.
b) Tome la muestra requerida en a); use un tipo de asignación proporcional y estime, junto con su error estándar de estimación, la proporción de proyectos de ley presentados por congresistas del partido nacionalista Gana Perú en la legislatura 2012-2013.

Capítulo 4

Muestreo por conglomerados

Los diseños muestrales estudiados presuponían la existencia de un marco muestral bien conocido y disponible, donde puede recabarse información que identifique a las posibles unidades seleccionadas de la población objetivo $\mathcal{P} = \{1, 2, ..., N\}$. En muchas situaciones este marco no se encuentra disponible y su elaboración puede ser muy costosa, e incluso imposible. En tales circunstancias, se puede dividir la población en grupos o agregados de estas unidades, y aplicar el muestreo sobre estos. Dichos grupos o agregados se denominan conglomerados.

Supongamos, por ejemplo, que deseamos hacer una encuesta de opinión en un distrito de la ciudad y que este distrito se encuentra dividido en barrios; es decir, en pequeñas zonas geográficas determinadas por calles, plazas, etc. En un muestreo por conglomerados se seleccionará primero una muestra de tales barrios y a continuación se averiguará la opinión de las personas en los barrios seleccionados. Esto último puede realizarse tomando en cuenta la opinión de todos los habitantes en los barrios seleccionados (muestreo por conglomerado de una etapa) o haciendo el estudio mediante nuevos muestreos al interior de cada barrio seleccionado (muestreo por conglomerados de dos o más etapas).

Como se ve, en el muestreo por conglomerados se parte de una estructura de subpoblaciones que conforman una partición de la población, como ocurre en el muestreo estratificado, pero la similitud termina aquí. En el muestreo estratificado se obtiene una muestra de cada estrato. En el muestreo por conglomerado se obtiene una muestra de conglomerados. Más áun, a fin de obtener diseños óptimos, la idea es que los conglomerados sean muy homogéneos entre sí y heterogéneos al interior, característica contraria a la de los estratos.

Entre las razones para optar por un muestreo de este tipo, citamos las siguientes:

- Es difícil, caro o imposible construir una lista de unidades de observación para el marco muestral.

- La población podría estar muy dispersa geográficamente o aparecer en cúmulos naturales como familias, centros penitenciarios, hospitales o escuelas.

Si bien, al contrario de un muestreo estratificado, un muestreo por conglomerados tiende a disminuir la precisión de las estimaciones con relación al muestreo aleatorio simple, el muestreo por conglomerados es el diseño más utilizado para encuestas sobre grandes poblaciones debido a su economía y sencillez en el trabajo de campo.

4.1. Teoría del muestreo por conglomerados

En un MAS, las unidades seleccionadas son los elementos observados. En un muestreo por conglomerados, las unidades seleccionadas o primarias son los conglomerados, y los elementos observados en su interior constituyen las unidades secundarias. En lo que resta de este capítulo utilizaremos las siguientes notaciones, donde abordaremos con detalle hasta el muestreo por conglomerados de dos etapas o bietápico. En primer lugar, nuestra población estará conformada por N unidades primarias (conglomerados) a los que llamaremos UPM (unidades primarias de muestreo). En cada UPM i asumiremos que existen M_i unidades secundarias, a las que llamaremos USM (unidades secundarias de muestreo), y será de interés estudiar una variable estadística y, cuyo valor para la j-ésima USM dentro de la UPM i será denotado por y_{ij}. Sean, por otro lado,

- $K = \sum_{i=1}^{N} M_i$ = número total de USM en la población
- n = número de UPM en la muestra
- m_i = número de USM en la muestra dentro de la UPM i

y definamos la variable aleatoria indicadora δ_{ij} como 1 si el elemento y_{ij} es seleccionado en la muestra, y 0 en caso contrario. Note que esta variable puede descomponerse como

$$\delta_{ij} = \delta_{j|i}\delta_i,$$

donde $\delta_{j|i}$ es una variable indicadora que toma el valor 1 si de seleccionarse la UPM i se selecciona en la segunda etapa la USM j; mientras que la variable indicadora δ_i vale 1 si, y solo si, la UPM i es seleccionada durante la primera etapa. En tal sentido, si las dos etapas se realizan mediante un MASs, podremos escribir la función de probabilidad de δ_{ij} mediante

$$P(\delta_{ij} = 1) = \frac{nm_i}{NM_i}.$$

En un muestreo de una etapa solo es necesario seleccionar las UPM, razón por la cual la variable δ_{ij} se reduce a δ_i, la cual vale 1 si la i-ésima UPM es seleccionada en la muestra, y 0 en caso contrario. Aquí se tiene que

$$P(\delta_i = 1) = \frac{n}{N}.$$

El siguiente cuadro muestra algunas de las cantidades de interés en nuestro estudio:

Denominación	Parámetro poblacional	Estimador puntual
Total de la UPM i	$\tau_i = \sum_{j=1}^{M_i} y_{ij}$	$\hat{\tau}_i = \frac{M_i}{m_i} \sum_{j=1}^{M_i} y_{ij}\delta_{j\|i}$
Media en la UPM i	$\mu_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$	$\bar{Y}_i = \frac{1}{m_i} \sum_{j=1}^{M_i} y_{ij}\delta_{j\|i}$
Media global	$\mu = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$	$\bar{Y} = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{NM_i}{nm_i} y_{ij}\delta_{ij}$
Varianza en la UPM i	$\sigma_i^2 = \frac{1}{M_i-1} \sum_{j=1}^{M_i} (y_{ij} - \mu_i)^2$	$S_i^2 = \frac{1}{m_i-1} \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2\delta_{j\|i}$
Varianza entre UPM	$\sigma_c^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\tau_i - \mu_\tau)^2$	$S_c^2 = \frac{1}{n-1} \sum_{i=1}^{N} (M_i\bar{Y}_i - \frac{K}{N}\bar{Y})^2\delta_i,$

Cuadro 4.1: Principales parámetros y estimadores puntuales en un muestreo por conglomerados bietápico

donde $\mu_\tau = \frac{1}{N} \sum_{i=1}^{N} \tau_i$ es la media de los totales de los conglomerados.

La razón de que el estimador puntual de la media poblacional tome una forma un tanto peculiar, es que este se construye con el fin de que sea un estimador insesgado de la media poblacional μ. En efecto, un simple cálculo muestra que

$$E(\bar{Y}) = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{NM_i}{nm_i} y_{ij} E(\delta_{ij}) = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{NM_i}{nm_i} y_{ij} \frac{nm_i}{NM_i} = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \mu.$$

Más adelante exploraremos otra forma de estimar μ.

4.2. Muestreo por conglomerados de una etapa

Como mencionamos, en el muestreo por conglomerados de una etapa se selecciona mediante un MASs una muestra de n UPM y se procede luego a medir la variable de interés en todos los elementos de los conglomerados seleccionados. Así, para el muestreo por conglomerados de una etapa, $m_i = M_i$. Con el fin de estimar la media en este diseño, se podrían considerar las medias de los conglomerados seleccionados, o funciones de ellas, como observaciones e ignorar los elementos individuales. El estimador insesgado de la media global μ para un muestreo por conglomerados de una etapa viene dado por

$$\bar{Y} = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{N}{n} y_{ij}\delta_i = \sum_{i=1}^{N} \frac{N}{nK} \tau_i\delta_i = \sum_{i=1}^{N} \frac{\tau_i}{n\overline{M}}\delta_i, \tag{4.1}$$

donde $\overline{M} = \frac{K}{N}$ es el tamaño promedio de los conglomerados. Note que esta expresión puede escribirse también como

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{N} (\frac{\tau_i}{\overline{M}})\delta_i,$$

lo cual sugiere qué se tomen como observaciones (agregadas) los elementos $\frac{\tau_i}{M}$. En consecuencia, por la teoría del MASs, la varianza de este estimador viene dada por

$$V(\bar{Y}) = (1 - \frac{n}{N})\frac{\sigma_m^2}{n},$$

donde $\sigma_m^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\frac{\tau_i}{M} - \mu)^2$. Como recordamos, un estimador insesgado de este último es la varianza muestral

$$S_m^2 = \frac{1}{n-1}\sum_{i=1}^{N}(\frac{\tau_i}{M} - \bar{Y})^2\delta_i.$$

Así, el error estándar de estimación estimado de la media $\bar{Y}$ resulta ser

$$\hat{SE}(\bar{Y}) = \sqrt{\hat{V}(\bar{Y})} = \sqrt{(1 - \frac{n}{N})\frac{S_m^2}{n}}.$$

Suponiendo que tenemos información de una muestra piloto o de un estudio anterior sobre S_m^2, este error de estimación nos permitirá, como es rutina en el MAS, obtener la siguiente fórmula para el tamaño de muestra de conglomerados para un nivel de confianza de $100(1-\alpha)\,\%$ y un error máximo de estimación para μ de e:

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 S_m^2 N}{z_{1-\frac{\alpha}{2}}^2 S_m^2 + e^2 N}.$$

Ejemplo 4.1. *Un problema con el estimador insesgado es que este posee en general una varianza grande, situación que se torna más crítica cuando los conglomerados son de distintos tamaños. Ello es natural, pues la varianza de este es la de los elementos $\frac{\tau_i}{M}$, los cuales poseen un denominador común, pero totales que varían mucho según la cantidad de unidades que contiene el conglomerado. Para ilustrar ello, retomemos la data de la ECE 2019, donde los conglomerados naturales en dicha población son los colegios. Supongamos ahora que deseamos estimar insesgadamente el rendimiento medio en Ciencia y Tecnología de los alumnos de la DRE Amazonas con un margen de error de no más de 5 puntos y una confianza del 95 % mediante un muestreo por conglomerados de una etapa. Si analizamos los totales de rendimiento de los colegios en esta población a través del histograma de la figura 4.1, vemos que, en efecto, este es altamente variable con un CV del 25.59 %*

```
load("ece19Am.RData")
tau_CT = ece19Am$M500_CT
```

Para calcular el tamaño de muestra (número de colegios) requeriremos estimar la varianza entre los elementos $\frac{\tau_i}{M}$ de los colegios. Si bien esta cantidad la podríamos obtener de un estudio piloto o un muestreo pasado, aquí la obtendremos solo como ilustración de nuestra data censal. El siguiente código nos permitirá realizar este cálculo

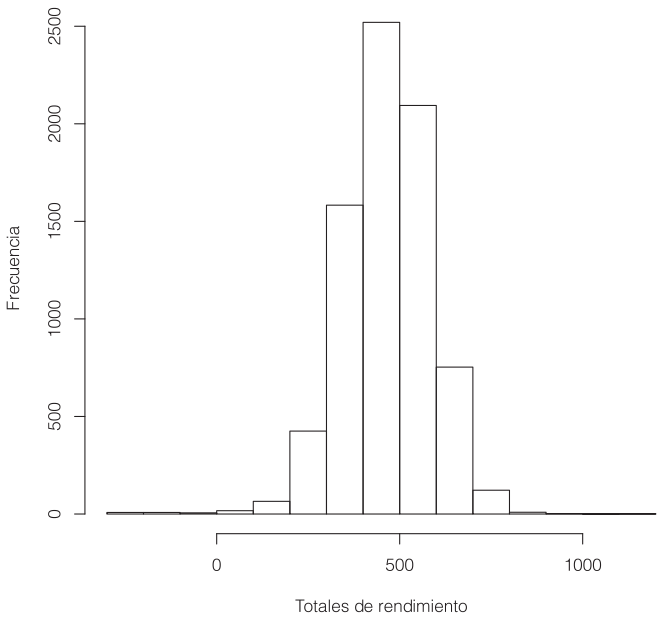


Figura 4.1: Histograma de los totales de rendimiento en Ciencia y Tecnología de los colegio de la DRE Amazonas en la ECE 2019

```
K = dim(ece19Am)[1]  #Número de estudiantes en Amazonas
(N = length(unique(ece19Am$ID_IE))) #Número de colegios en Amazonas

## [1] 286

Mbar = K/N
aux = aggregate(ece19Am$M500_CT,by=list(ece19Am$ID_IE),sum)
Sm2 = var(aux$x/Mbar,na.rm=T)
```

El número de colegios a seleccionarse será entonces:

```
d2 = 25*N/(qnorm(0.975)^2)
ceiling(Sm2*N/(d2 + Sm2))

## [1] 281
```

que, como se aprecia, es sumamente alto e implica casi un censo. □

4.3. El estimador de razón

En la sección anterior indirectamente hemos asumido que K o $\overline{M}$ eran cantidades conocidas. Usualmente, sin embargo, uno desconoce los tamaños de todos los conglomerados. Si retomamos el estimador insesgado (4.1) de μ,

$$\bar{Y} = \sum_{i=1}^{N} \frac{\tau_i}{n\overline{M}} \delta_i,$$

vemos que una idea para salvar tal problema podría consistir en estimar $\overline{M}$ como el tamaño promedio de solo los conglomerados seleccionados en la muestra; vale decir, por

$$\frac{1}{n} \sum_{i=1}^{N} M_i \delta_i.$$

Si hacemos esto, obtendremos entonces el llamado estimador de razón de μ:

$$\bar{Y}_r = \frac{\sum_{i=1}^{N} \tau_i \delta_i}{\sum_{i=1}^{N} M_i \delta_i}.$$

Este, como su nombre lo sugiere, es un cociente o una razón entre dos variables aleatorias. Si bien el estudio teórico de este tipo de estimadores lo haremos en el capítulo 5, adelantaremos la siguiente proposición de importancia práctica para este estimador.

Proposición 4.1. *El error estándar de estimación aproximado para el estimador de razón viene dado por*

$$SE(\bar{Y}_r) = \sqrt{(1 - \frac{n}{N}) \frac{1}{n(n-1)\bar{M}^2} \sum_{i=1}^{N} M_i^2 (\bar{Y}_i - \bar{Y}_r)^2 \delta_i},$$

donde de ser $\bar{M}$ desconocido, este puede estimarse por $\frac{1}{n} \sum_{i=1}^{N} M_i \delta_i$.

Al igual que en los diseños anteriores, podemos utilizar el último resultado para obtener un tamaño de muestra de conglomerados que nos permita estimar μ con un error máximo e y un nivel de confianza del $100(1-\alpha)\,\%$. Ello se obtiene de despejar n en la ecuación

$$e = z_{1-\frac{\alpha}{2}} SE(\bar{Y}_r) = z_{1-\frac{\alpha}{2}} \sqrt{(1 - \frac{n}{N}) \frac{1}{n\bar{M}^2} S_r^2},$$

donde:

$$S_r^2 = \frac{1}{n-1} \sum_{i=1}^{N} M_i^2 (\bar{Y}_i - \bar{Y}_r)^2 \delta_i = \frac{1}{n-1} \sum_{i=1}^{N} (\hat{\tau}_i - M_i \bar{Y}_r)^2 \delta_i$$

ha de estimarse de una prueba piloto o de un estudio similar, y $\bar{M}$, de ser desconocido, puede estimarse por $\frac{1}{n}\sum_{i=1}^{N} M_i\delta_i$. Un poco de álgebra nos lleva entonces a la siguiente fórmula:

$$n = \frac{NS_r^2}{N(\frac{e\bar{M}}{z_{1-\frac{\alpha}{2}}})^2 + S_r^2}.$$

Cabe indicar, sin embargo, que esta fórmula es válida para tamaños de muestra suficientemente grandes. Ello lo valida el teorema del límite central al construirse un intervalo de confianza aproximado para μ. El sesgo del estimador de razón se hace cada vez más despreciable conforme aumenta n, y la varianza de este estimador resulta ser por lo común mucho menor que la del estimador insesgado, en especial si los tamaños de los conglomerados muestran una alta heterogeneidad.

Observación: Si los tamaños de los conglomerados son todos iguales, entonces el estimador insesgado y de razón para μ coinciden.

4.4. Estimación de una proporción

Si recordamos que una proporción no es sino la media de una variable dicotómica Y, entonces todo el análisis anterior es completamente similar si trabajamos ahora con una variable de tal tipo. En este caso, los estimadores puntuales de la proporción p de elementos de la población que comparten una característica dada para la cual Y vale 1 vienen dados por

$$\hat{p} = \sum_{i=1}^{N} \frac{a_i}{n\bar{M}}\delta_i$$

en el caso insesgado y por

$$\hat{p}_r = \frac{\sum_{i=1}^{N} a_i\delta_i}{\sum_{i=1}^{N} M_i\delta_i}$$

para el estimador de razón, siendo a_i el número de elementos en el conglomerado i que comparten la característica dada. Todas las demás propiedades de la media se verifican para la proporción al reemplazar τ_i por a_i.

Ejemplo 4.2. *Un sociólogo desea estimar los ingresos anuales medios por persona de cierta ciudad, así como la proporción de estas personas que alquilan sus viviendas (es decir, que no son propietarios). Dado que él no dispone de una lista de las personas adultas residentes, decide tomar una muestra por conglomerados. Para ello, obtiene un mapa de la ciudad que lo divide en 415 bloques rectangulares. Luego selecciona al azar 25 de ellos y asigna un grupo de encuestadores a cada uno de los conglomerados seleccionados para que recaben la información requerida en todos los hogares de dichos conglomerados. Luego del trabajo de campo se obtuvo la tabla que seguidamente se detalla, donde los ingresos están en cientos de dólares.*

a) *Estime puntualmente la proporción de arrendatarios en esta ciudad y establezca un límite para el error de estimación con una confianza del 95 %.*

b) *Si era de interés para el sociólogo estimar el ingreso anual medio por persona en esta ciudad con un error máximo de 100 dólares, ¿fue suficiente el tamaño de muestra tomado?*

Conglomerado	*Número de residentes adultos*	*Ingresos totales*	*Número de personas que alquilan*
1	*8*	*96*	*4*
2	*12*	*121*	*7*
3	*4*	*42*	*1*
4	*5*	*65*	*3*
5	*6*	*52*	*3*
6	*6*	*40*	*4*
7	*7*	*75*	*4*
8	*5*	*65*	*2*
9	*8*	*45*	*3*
10	*3*	*50*	*2*
11	*2*	*85*	*1*
12	*6*	*43*	*3*
13	*5*	*54*	*2*
14	*10*	*49*	*5*
15	*9*	*53*	*4*
16	*3*	*50*	*1*
17	*6*	*32*	*4*
18	*5*	*22*	*2*
19	*5*	*45*	*3*
20	*4*	*37*	*1*
21	*6*	*51*	*3*
22	*8*	*30*	*3*
23	*7*	*39*	*4*
24	*3*	*47*	*0*
25	*8*	*41*	*3*

Solución: **a)** *Puesto que no conocemos aquí el tamaño de los conglomerados no seleccionados, solo podríamos usar el estimador de razón. Este y su error de estimación lo podríamos calcular introduciendo los datos en la base de datos Rentas y utilizando el código*

```
N = 415
n = 25
load('Rentas.RData')
Mi = Rentas$Nresidentes
pi = Rentas$Nalquilan/Mi
(pr = sum(Rentas$Nalquilan)/sum(Mi))
```

```
## [1] 0.477

S2pr = sum(Mi^2*(pi-pr)^2)/(n-1)
SEpr_e = sqrt((1-n/N)*S2pr/(n*mean(Mi)^2))
(e = qnorm(0.975)*SEpr_e)

## [1] 0.0458
```

b) *Para responder a esto podríamos hallar el error de estimación máximo con el tamaño actual de muestra o el tamaño de muestra para* $e = 1$. *Optemos por el segundo camino. Este tamaño de muestra debería ser de*

```
Ybarr = sum(Rentas$Ingresos_totales)/sum(Mi)
Ybari = Rentas$Ingresos_totales/Mi
S2r = sum((Mi*(Ybari-Ybarr))^2)/(n-1)
D = mean(Mi)/qnorm(0.975)
(n1= ceiling(N*S2r/(N*D^2 + S2r)))

## [1] 58
```

por lo que el tamaño de muestra tomado no fue suficiente. □

4.5. Muestreo por conglomerado bietápico

En el muestreo por conglomerados de una etapa examinamos todas las USM dentro de cada UPM elegida. En muchas situaciones, sin embargo, los conglomerados pueden ser demasiado similares o numerosos, de modo que el análisis de todas las subunidades dentro de una unidad primaria será un desperdicio de recursos. En estos casos podría ser más eficiente y económico tomar una submuestra dentro de cada UPM. Este muestreo se denomina un muestreo por conglomerados bietápico y se resume como sigue:

- Se considera un MASs de tamaño n sobre la población de N UPM.
- Se considera un MASs de m_i USM dentro de cada UPM i seleccionada.

Como vimos en la tabla 4.1, un estimador insesgado de la media global poblacional μ viene dado por

$$\bar{Y} = \frac{1}{K}\sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{NM_i}{nm_i}y_{ij}\delta_{ij}.$$

Dado que ahora se toman dos muestras, la varianza de este estimador posee dos componentes, una debido a la variabilidad entre las UPM y otra debido a la variabilidad entre las USM al

interior de cada UPM. Concretamente, se puede probar (más adelante mostraremos un caso más general) que

$$V(\bar{Y}) = (1 - \frac{n}{N})\frac{\sigma_c^2}{n\bar{M}^2} + \frac{1}{n\bar{M}^2 N}\sum_{i=1}^{N}(1 - \frac{m_i}{M_i})M_i^2\frac{\sigma_i^2}{m_i}, \tag{4.2}$$

donde σ_c^2 es la varianza entre UPM y σ_i^2 es la varianza al interior de la UPM i. Estas últimas cantidades se pueden estimar, respectivamente, por

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{N}(\hat{\tau}_i - \bar{M}\bar{Y})^2\delta_i$$

y S_i^2, dando lugar al siguiente estimador insesgado de la varianza (4.2):

$$\hat{V}(\bar{Y}) = (1 - \frac{n}{N})\frac{S_c^2}{n\bar{M}^2} + \frac{1}{n\bar{M}^2 N}\sum_{i=1}^{N}(1 - \frac{m_i}{M_i})M_i^2\frac{S_i^2}{m_i}\delta_i.$$

Al igual que en el muestreo por conglomerados de una etapa, el problema con $\bar{Y}$ recae en el desconocimiento de K. Una manera de subsanar ello es utilizando nuevamente el estimador de razón

$$\bar{Y}_r = \frac{\sum_{i=1}^{N}\hat{\tau}_i\delta_i}{\sum_{i=1}^{N}M_i\delta_i} = \frac{\sum_{i=1}^{N}M_i\bar{Y}_i\delta_i}{\sum_{i=1}^{N}M_i\delta_i}.$$

Si bien este es un estimador sesgado, tal sesgo es despreciable para n grande y usualmente este estimador posee una varianza menor que la del estimador insesgado. Esta última se prueba que viene dada aproximadamente por

$$V(\bar{Y}_r) = (1 - \frac{n}{N})\frac{\sigma_r^2}{n\bar{M}^2} + \frac{1}{n\bar{M}^2 N}\sum_{i=1}^{N}(1 - \frac{m_i}{M_i})M_i^2\frac{\sigma_i^2}{m_i},$$

donde:

$$\sigma_r^2 = \frac{1}{N-1}\sum_{i=1}^{N}M_i^2(\mu_i - \mu)^2 = \frac{1}{N-1}\sum_{i=1}^{N}(M_i\mu_i - M_i\mu)^2.$$

Un estimador de esta última cantidad es

$$\hat{V}(\bar{Y}_r) = (1 - \frac{n}{N})\frac{S_r^2}{n\bar{M}^2} + \frac{1}{n\bar{M}^2 N}\sum_{i=1}^{N}(1 - \frac{m_i}{M_i})M_i^2\frac{S_i^2}{m_i}\delta_i,$$

donde:

$$S_r^2 = \frac{1}{n-1}\sum_{i=1}^{N}M_i^2(\bar{Y}_i - \bar{Y}_r)^2\delta_i = \frac{1}{n-1}\sum_{i=1}^{N}(\hat{\tau}_i - M_i\bar{Y}_r)^2\delta_i.$$

Para estimar una proporción basta recordar que este es un caso particular de estimación de la media cuando la variable de investigación es dicotómica. Luego, uno puede fácilmente

encontrar que el estimador de razón de la proporción de elementos de la población p para los cuales la variable toma el valor 1 viene dado por

$$\hat{p}_r = \frac{\sum_{i=1}^{N} M_i \hat{p}_i \delta_i}{\sum_{i=1}^{N} M_i \delta_i},$$

siendo $\hat{p}_i$ la proporción estimada en la muestra del conglomerado i. La varianza estimada de este estimador viene dada por

$$\hat{V}(\hat{p}_r) = (1 - \frac{n}{N})\frac{S_r^2}{n\bar{M}^2} + \frac{1}{n\bar{M}^2 N}\sum_{i=1}^{N}(1 - \frac{m_i}{M_i})M_i^2 \frac{\hat{p}_i(1-\hat{p}_i)}{m_i - 1},$$

donde:

$$S_r^2 = \frac{1}{n-1}\sum_{i=1}^{N} M_i^2(\hat{p}_i - \hat{p}_r)^2\delta_i = \frac{1}{n-1}\sum_{i=1}^{N}(M_i\hat{p}_i - M_i\hat{p}_r)^2\delta_i.$$

4.6. La correlación intraclase y el efecto de diseño

Esta sección se enfoca en comparar un muestreo por conglomerados con un MAS, para lo cual será de gran utilidad introducir alguna medida que refleje la variabilidad existente al interior de los conglomerados con relación a la variabilidad de toda la población. Para tal efecto será de gran ayuda analizar la siguiente descomposición de esta última variabilidad, la cual podría medirse por la suma de cuadrados totales $SCT = \sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij}-\mu)^2$. Sumando y restando en el término cuadrático la media μ_i de cada conglomerado, obtendremos que

$$\overbrace{\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij}-\mu)^2}^{SCT} = \overbrace{\sum_{i=1}^{N} M_i(\mu_i-\mu)^2}^{SCC} + \overbrace{\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij}-\mu_i)^2}^{SCE},$$

donde a SCC se le denomina la suma de cuadrados entre conglomerados; y a SCE, la suma de cuadrados del error o dentro de los conglomerados. Esta descomposición puede también resumirse en la tabla ANOVA del cuadro 4.2.

Fuente de variabilidad	Sumas de cuadrados	Número de términos
Entre conglomerados	SCC	N
Dentro de los conglomerados	SCE	$K - N$
Total	SCT	K

Cuadro 4.2: Tabla ANOVA para un muestreo por conglomerados.

Sobre la base del último cuadro, una posible medida de homogeneidad al interior de los conglomerados viene dada por el coeficiente de determinación ajustado

$$R_a^2 = 1 - (\frac{K}{K-N})\frac{SCE}{SCT}.$$

Mientras más cercano esté R_a^2 a 1, más homogéneos serán los conglomerados y la variabilidad existente será explicada en esencia por las diferencias entre las medias de los conglomerados.

Un caso particular del análisis anterior se da cuando los tamaños M_i de los conglomerados son todos iguales, digamos M. En este caso, al coeficiente

$$R_a^2 = 1 - (\frac{M}{M-1})\frac{SCE}{SCT}$$

se le suele denotar por ρ y se le denomina la correlación intraclase. Una de las razones de su popularidad es que se puede probar que ρ no es sino el coeficiente de correlación de Pearson para los $NM(M-1)$ distintos posibles pares (y_{ij}, y_{ik}), con $i = 1, 2, \ldots, N$ y $j \neq k \in \{1, 2, \ldots, M\}$, que uno pudiera tomar en la población de y (véase el ejercicio 4.7).

Veamos ahora el rol que desempeña ρ en el cálculo del efecto de diseño para un muestreo por conglomerados de una etapa cuando los conglomerados tienen el mismo tamaño. Como recordamos, para este último caso, la varianza del estimador de la media poblacional viene dada por

$$V_c(\bar{Y}) = (1 - \frac{n}{N})\frac{1}{nM(N-1)}SCC;$$

mientras que la varianza de este estimador bajo un MASs es

$$V_{MASs}(\bar{Y}) = (1 - \frac{n}{N})\frac{\sigma^2}{nM} = (1 - \frac{n}{N})\frac{SCT}{nM(NM-1)}.$$

Relacionando SCC con ρ, obtenemos

$$\frac{SCC}{SCT} = 1 - \frac{SCE}{SCT} = 1 - \frac{M-1}{M}(1-\rho) = \frac{1+\rho(M-1)}{M}$$

y, consecuentemente:

$$SCC = SCT\frac{1+\rho(M-1)}{M}.$$

Reemplazando esta expresión en la varianza V_c y tomando el cociente con la varianza V_{MASs}, resulta que el efecto de diseño viene dado por

$$deff = \frac{V_c(\bar{Y})}{V_{MASs}(\bar{Y})} = \frac{NM-1}{M(N-1)}(1+\rho(M-1)). \tag{4.3}$$

Dado que $\frac{NM-1}{M(N-1)} > 1$, este efecto será mayor que 1 y, por tanto, el diseño por conglomerados de una etapa será menos eficiente que el MASs si $\rho > 0$. Esta es, en efecto, la situación más usual. Aquí, los elementos al interior de los conglomerados tienden a ser más similares entre sí que los elementos seleccionados aleatoriamente de la población, lo cual básicamente ocurre porque los elementos al interior de un conglomerado comparten un entorno similar; así en el caso de una encuesta de hogares se esperará que los miembros de una vecindad seleccionada (conglomerado), que han optado por vivir en ella y a interactuar con sus vecinos, tiendan a compartir varias características comunes o posean opiniones similares ante distintos cuestionamientos.

Raramente el efecto de diseño será menor que 1, y esto sucederá cuando $\rho < -\frac{1}{NM-1}$.

Ejemplo 4.3. *Consideremos nuevamente el ECE 2019 para la DRE Amazonas y calculemos para ejemplificar, pues tenemos a toda la población, el coeficiente de determinación ajustado en la estimación de los rendimientos de Matemáticas. Este viene dado por*

```
fit = aov(ece19Am$M500_M ~ factor(ece19Am$ID_IE),data=ece19Am)
(R2a =1-(1-1/K)*summary(fit)[[1]]$'Mean Sq'[2]/var(ece19Am$M500_M,na.rm=T))

## [1] 0.49
```

Como se aprecia, la prueba ANOVA, que resulta significativa, tiene un coeficiente de determinación ajustado alto y positivo. □

4.7. Muestreo sistemático

Considere una población con N elementos, donde por simplicidad supondremos que $N = nk$, siendo k un número natural y n el tamaño de muestra a considerar. Asumamos también que disponemos de un marco muestral ordenado: $1, 2, \ldots, N$. Si seleccionamos ahora al azar una unidad de entre los primeros k, digamos la unidad j, y luego consecutivamente los siguientes $n-1$ elementos tomados de k en k; es decir, los elementos

$$j+k,\ j+2k,\ \ldots,\ j+(n-1)k,$$

entonces diremos que hemos empleado en esta selección un muestreo sistemático.

La principal ventaja de un muestreo sistemático es su sencillez de ejecución. También es'a sujeto a menos posibilidades de errores por parte del entrevistador. En cuanto a su precisión, esta depende de la muestra y no es posible su directa evaluación con ella. Para entender ello resulta revelador considerar el muestreo sistemático como un caso particular de un muestreo por conglomerados. En efecto, si escribimos los valores de nuestra variable estadística de interés en la población como

$$y_1,\ \ldots,\ y_k,\ y_{k+1},\ \ldots,\ y_{2k},\ y_{2k+1},\ \ldots,\ y_{(n-1)k},\ y_{(n-1)k+1},\ \ldots, y_{nk}$$

o mejor aún en una matriz como

Muestras	1	2	...	n	Medias
1	y_1	y_{k+1}	$\cdots$	$y_{(n-1)k+1}$	μ_1
2	y_2	y_{k+2}	$\cdots$	$y_{(n-1)k+2}$	μ_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_i	y_{k+i}	$\cdots$	$y_{(n-1)k+i}$	μ_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	y_k	y_{2k}	$\cdots$	y_{nk}	μ_k

(4.4)

Cada fila de esta matriz representa una posible muestra sistemática de tamaño n, con su respectiva media. Por tanto, podríamos considerar estas filas conglomerados de igual tamaño, y de los cuales seleccionamos tan solo uno. Así, el muestreo sistemático se reduce a un muestreo por conglomerados de una etapa con k conglomerados, cada uno de tamaño n, de donde seleccionamos tan solo uno. Consecuentemente, el estimador insesgado de

$$\mu = \frac{1}{nk}\sum_{i=1}^{nk} y_i$$

viene dado por

$$\hat{\mu} = \bar{Y}_\alpha,$$

siendo α el conglomerado seleccionado. Más aún, la varianza de este estimador viene dada por

$$Var(\hat{\mu}) = \frac{1}{k}\sum_{i=1}^{k}(\mu_i - \mu)^2, \tag{4.5}$$

y el efecto de diseño al estimar la media en un muestreo sistemático toma la forma

$$deff = \frac{N-1}{N-n}(1 + \rho(n-1)).$$

Vemos entonces que si ρ es cercano a 1, los elementos dentro de la muestra serán bastante similares con respecto a la característica que estamos midiendo, y el muestreo sistemático producirá una varianza de la media muestral mayor que la obtenida con un MASs. Si ρ es negativo, entonces el muestreo sistemático puede ser más preciso que el MAS. La correlación puede ser negativa si los elementos dentro de la muestra sistemática tienden a ser extremadamente diferentes. Para ρ cercano a 0 y N bastante grande, el muestro sistemático es aproximadamente equivalente al MASs.

Ejemplo 4.4. *Consideremos la siguiente población de un centro de trabajo:*

Sujeto	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Edad	*35*	*24*	*60*	*38*	*22*	*33*	*54*	*45*	*38*	*19*	*53*	*40*
Sexo	*M*	*H*	*H*	*M*	*H*	*M*	*M*	*M*	*H*	*M*	*M*	*H*
Ingreso	*3333*	*3401*	*7687*	*3531*	*3134*	*3087*	*4813*	*4113*	*5064*	*2017*	*4724*	*5300*

donde es de interés estimar el ingreso medio de estas personas sobre la base de una muestra sistemática de tamaño $n = 4$. Obtenga esta estimación y calcule el efecto de este diseño.
Solución: Definamos inicialmente en R nuestra población de estudio.

```
Popc=data.frame(Sujeto=c(1:12),Edad = c(35,24,60,38,22,33,54,45,38,19,53,40),
Sexo = c("M","H","H","M","H","M","M","M","H","M","M","H"),
Ingreso = c(3333,3401,7687,3531,3134,3087,4813, 4113,5064,2017,4724,5300))
```

Para obtener la muestra sistemática deberíamos ordenar los datos en una matriz 3×4, seleccionar un número aleatorio entre 1 y 3, tomar la fila obtenida y, finalmente, obtener la estimación pedida, como el promedio de los datos en esa fila. Esto en R se hace con

```
set.seed(12345)
M = matrix(Popc$Ingreso,nrow=3,ncol=4)
m = sample(3,1)
MuestraS = data.frame(Ingreso = M[m,])
(Media = mean(MuestraS$Ingreso))

## [1] 5284
```

Dado que en este ejemplo disponemos de toda la población, es factible obtener la correlación intraclase. Esta por definición es

```
Popc = cbind(Popc,cluster=rep(1:3,4))
N = dim(Popc)[1]
n = 4
fit = aov(Popc$Ingreso~factor(Popc$cluster),data=Popc)
SCE = summary(fit)[[1]]$'Sum Sq'[2]
SCT = summary(fit)[[1]]$'Sum Sq'[1] + SCE
(rho1 = 1 - (n/(n-1))*(SCE/SCT))

## [1] 0.0946
```

Desde otro punto de vista, podríamos también calcular la correlación intraclase usando el paquete `combinat` *mediante*

```
library(combinat)
k = max(Popc$cluster)
gx <-function(x,r){c(M[r,x[1]],M[r,x[2]])}
pairs = cbind(combn(1:4,2,gx,simplify=T,1), combn(4:1,2,gx,simplify=T,1))
for (j in 2:k){
pairs = cbind(pairs,cbind(combn(1:4,2,gx,simplify=T,j),
combn(4:1,2,gx,simplify=T,j)))}
(rho2 = cor(t(pairs))[1,2])

## [1] 0.0946

(deff = (N-1)*(1 + rho2*(n-1))/(N-n))

## [1] 1.77
```

Como se aprecia, se obtiene una correlación intraclase cercana a 0 y un efecto de diseño de aproximadamente 1.765, lo que nos da un diseño un poco menos preciso que el MASs. □

Un problema central con el muestreo sistemático es, como adelantamos, que este no nos permite obtener una estimación directa de la varianza del estimador, ya que solo se basa en una muestra de un único conglomerado. Una solución podría ser considerar la fórmula de un MASs, lo cual para los datos del ejemplo anterior podría ser algo razonable. Si hiciéramos eso, el error estándar de estimación estimado sería

```
library(survey)
disC = svydesign(id=~1,fpc=rep(12,4),data=MuestraS)
svymean(~Ingreso,disC)

##         mean  SE
## Ingreso 5284 769
```

el cual subestima ligeramente al verdadero error estándar de estimación de la media muestral. En efecto, la media y el error estándar de la media muestral (valores que conocemos, solo porque tenemos a toda la población) vienen dados por

```
c(mean(Popc$Ingreso),sqrt(deff*(1-n/N)*var(Popc$Ingreso)/n))

## [1] 4184  797
```

Todo el análisis previo se realizó partiendo de un determinado orden para el marco muestral. Si este orden cambia, las estimaciones ciertamente también lo harán, por lo cual es importante conocer algo de la estructura de la población. En tales circunstancias, y dada la falta de información sobre esta, se recomienda que el diseño sea asistido bajo un modelo. En efecto, nosotros podríamos modelar el orden de la generación de datos en el marco muestral bajo principalmente tres asunciones:

- El marco muestral está en un orden aleatorio y no tiene relación con la variable de interés. Esto es lo que asumimos en el ejemplo. En tal caso, la correlación intraclase resultará ser cercana a 0 y el diseño será muy similar a un MASs.

- El marco muestral podría estar ordenado en orden creciente o decreciente según la variable de interés o alguna variable relacionada. En tales casos de autocorrelación positiva, el muestreo sistemático resultará ser por lo general más preciso que un MASs al producir correlaciones intraclase negativas. Así, si usaramos un MASs para estimar el error estándar de estimación, estaremos usualmente sobreestimando esta cantidad. Como lo señalan Lehtonen y Pahkinen (2004), si $\rho_q > 0$ es el coeficiente de autocorrelación entre un par de unidades que están q unidades aparte, una mejor formulación

para el error estándar se obtendría con la formulación

$$\hat{SE}_q = \sqrt{(1-\frac{n}{N})\frac{S^2}{n}(1+\frac{2}{\log(\hat{\rho}_q)}+\frac{2\hat{\rho}_q}{1+\hat{\rho}_q})},$$

siendo $0 < \hat{\rho}_q < 1$ el valor estimado de la autocorrelación.

- El marco muestral presenta un patrón periódico. En tal caso, si seleccionamos las unidades coincidentemente con el mismo período, el muestreo sistemático será mucho menos preciso que el MASs.

Otras maneras de solucionar el problema anterior, es aplicando una estratificación implícita o realizando un muestreo sistemático replicado. El primero consiste en ordenar, en primer lugar, el marco muestral según la variable de interés o alguna relacionada (pues, la de interés se desconoce en la población). Esto determinará secuencialmente de manera implícita dos o más estratos, por lo cual la estimación del error estándar de estimación podrá obtenerse como si este fuera un MAE con asignación proporcional.

En el muestreo replicado, por otro lado, uno selecciona más de una muestra sistemática. Por ejemplo, 10 muestras sistemáticas con $k = 50$, conteniendo cada una 6 mediciones, podrían obtenerse en aproximadamente el mismo tiempo que una muestra sistemática con $k = 5$, conteniendo 60 mediciones. Ambos procedimientos generan 60 mediciones, pero solo el muestreo replicado permite estimar la varianza (4.5) utilizándose para ello la varianza de las 10 medias muestrales obtenidas. El promedio de las 10 medias muestrales estimaría la media poblacional μ.

Como ilustración de estas técnicas, retomemos nuevamente el ejemplo previo. Para la estratificación implícita utilizaremos la edad, que es una variable relacionada al ingreso (asumiendo que conocemos la edad de todos) y conformaremos dos pseudoestratos.

```
library(sampling)
Popco=Popc[order(Popc$Edad),]
Popco = cbind(Popco,Estimp=c(rep(1,6),rep(2,6)),fpc=rep(6,12))
m=strata(Popco,c("Estimp"),size=rep(2,2),method="srswor")
Popcosample=getdata(Popco,m)
DisEI = svydesign(ids=~1,stratum=~Estimp,fpc=~fpc,data=Popcosample)
svymean(~Ingreso,DisEI)

##         mean  SE
## Ingreso 4604 606
```

Por otro lado, para el muestreo replicado podríamos considerar 6 conglomerados de 2 observaciones cada uno y seleccionar al azar y sin reeemplazamiento a 2 de ellos, de tal manera que con las medias de estos podamos estimar el error estándar de estimación.

```
Popc = cbind(Popc,cluster1 = rep(1:6,2))
set.seed(12345)
s = sample(6,2)
MuestraR = Popc[Popc$cluster1 %in% s,]
(mR = mean(MuestraR$Ingreso))

## [1] 4061

mRc = as.vector(by(MuestraR$Ingreso,MuestraR$cluster1,mean))
(SER = sqrt(var(mRc)))

## [1] 187
```

4.8. Tamaños de muestra para diseños multietápicos

La elección de tamaños de muestra para un muestreo multietápico reviste gran complejidad, pues no solo es necesario determinar cuántas UPM seleccionar, sino también cuántas USM u otras unidades si hay más etapas. Veamos el caso de la estimación de una media poblacional bajo un muestreo bietápico, y para simplificar asumamos que las UPM son todas de igual tamaño M y se tomará una misma cantidad m de USM por cada UPM. Aparte de tomar en cuenta la precisión, será conveniente también incluir los costos de muestreo, los cuales se buscarán minimizar o prefijar. El costo total de muestreo lo asumiremos lineal y vendrá dado por $C = c_0 + c_1 n + c_2 nm$, donde c_0 es un costo fijo; c_1 es el costo unitario por cada UPM seleccionada, y c_2 es el costo unitario por cada USM seleccionada. En cuanto a la precisión, recordemos que esta se mide según (4.2) por

$$V(\bar{Y}) = (1 - \frac{n}{N})\frac{\sigma_c^2}{nM^2} + \frac{1}{nmN}(1 - \frac{m}{M})\sum_{i=1}^{N}\sigma_i^2,$$

donde cabe notar que la varianza entre UPM se relaciona con la suma de cuadrados entre conglomerados, definida en la sección 4.6, mediante $\sigma_c^2 = \frac{M}{N-1}SCC$, y la suma de las varianzas al interior de las UPM satisface $\sum_{i=1}^{N}\sigma_i^2 = \frac{SCE}{M-1}$. Para simplificar podrímos introducir, como se hace usualmente en el análisis de varianza, la media cuadrática entre conglomerados $MCC = \frac{SCC}{N-1}$ y la media cuadrática del error $MCE = \frac{SCE}{K-N} = \frac{SCE}{N(M-1)}$. Así, las relaciones anteriores nos dicen que $\sigma_c^2 = M \times MCC$ y $\sum_{i=1}^{N}\sigma_i^2 = N \times MCE$ y, en consecuencia,

$$\begin{aligned} V(\bar{Y}) &= (1 - \frac{n}{N})\frac{MCC}{nM} + \frac{1}{nm}(1 - \frac{m}{M})MCE \\ &= \frac{MCC - MCE}{nM} + \frac{MCE}{nm} - \frac{MCC}{NM}. \end{aligned}$$

Con el fin de determinar los tamaños de muestra n y m óptimos, usaremos como criterio minimizar la varianza anterior para un costo fijo total C o minimizar el costo total de muestreo para un valor fijo de la varianza anterior. Esto, como recordamos, puede resolverse de manera similar a lo visto en la demostración de la proposición 3.2; es decir, usándose la desigualdad de Cauchy-Schwartz que busca minimizar

$$(\frac{MCC - MCE}{nM} + \frac{MCE}{nm})(c_1 n + c_2 nm) = (\frac{MCC - MCE}{M} + \frac{MCE}{m})(c_1 + c_2 m).$$

Ello nos conduce a las siguientes formulaciones de tamaños de muestra óptimos:

$$m = \sqrt{\frac{M c_1 MCE}{c_2(MCC - MCE)}}$$

y

$$n = \frac{C - c_0}{c_1 + c_2 m}.$$

Expresando las medias cuadráticas en términos del coeficiente de correlación intraclase por $MCE = (1-\rho)\frac{SCT}{NM}$ y $MCC = (\frac{1+(M-1)\rho}{M(N-1)})SCT$, estas formulaciones podrían escribirse también como

$$m = \sqrt{\frac{M(N-1)(1-\rho)c_1}{(1+(NM-1)\rho)c_2}}$$

y

$$n = \frac{C - c_0}{c_1 + c_2 m}.$$

Note que si el número de conglomerados es suficientemente grande, se tendrá la aproximación

$$m = \sqrt{\frac{(1-\rho)c_1}{\rho c_2}},$$

y así la elección dependerá tan solo del costo relativo unitario y del coeficiente de correlación intraclase.

Un desarrollo similar se da, por ejemplo, para un muestreo trietápico. Véase el ejercicio 9 de este capítulo.

Cabe precisar que el tratamiento anterior es en parte elegante por el hecho de que se ha asumido que los conglomerados son de igual tamaño. Si ello no es así, uno tendrá en general que resolver numéricamente un problema de optimización. Sin embargo, como lo demuestran varios autores entre los que destacan Khan y Ahmad (2006), es posible flexibilizar algunas restricciones a fin de derivar una fórmula cerrada para los tamaños de muestra buscados. Detalles sobre cómo hacer esto se tienen en el artículo citado. Aquí mencionaremos solo los tamaños de muestra n y m_i para la UPM y las USM dentro de cada UPM i óptimos que

minimizan la varianza de $\bar{Y}$ bajo una restricción presupuestal de C_0 unidades monetarias en los costos de muestreo variables. Estos valores vienen dados por

$$n = \frac{C_0\sqrt{A}}{c_1\sqrt{A} + \sqrt{c_1 c_2}\sum_{i=1}^{N}\frac{M_i}{N}\sigma_i}$$

y

$$m_i = M_i\sigma_i\sqrt{\frac{c_1}{Ac_2}},$$

donde $A = \sigma_c^2 - \frac{1}{N}\sum_{i=1}^{N} M_i\sigma_i^2$.

4.9. El estimador de Horvitz-Thompson

Hasta el momento hemos asumido que las probabilidades de selección de primera etapa

$$\pi_i = P(\delta_i = 1)$$

son todas iguales e independientes de la unidad primaria considerada. Para ser más explícitos, en un muestreo por conglomerados bietápico estas estaban dadas por $\pi_i = \frac{n}{N}$, siendo N el número de UPM en la población y n el tamaño de muestra de UPM. La asunción de probabilidades iguales no siempre es la adecuada para algunos requerimientos. Un típico ejemplo es el de un muestreo por conglomerados con probabilidades proporcionales al tamaño (ppt). En este se exige que los conglomerados más grandes tengan mayores probabilidades de selección. Asumiendo, como lo hemos venido haciendo, un muestreo sin reemplazamiento, la selección de las unidades de la segunda etapa o posteriores se complica bajo este esquema, dado que ellas dependen de las unidades particulares seleccionadas en la primera etapa. Horvitz y Thompson (1952) propusieron que de obtenerse estimaciones insesgadas de los totales en cada unidad primaria, uno podría estimar el total de la población mediante

$$\hat{\tau}_{HT} = \sum_{i=1}^{N}\frac{\hat{\tau}_i}{\pi_i}\delta_i,$$

siendo $\hat{\tau}_i$ un estimador insesgado del total poblacional τ_i para la i-ésima UPM, el cual se asume que es independiente de δ_i. Como seguidamente se aprecia, este es un estimador insesgado del total poblacional τ,

$$E(\hat{\tau}_{HT}) = \sum_{i=1}^{N} E(\frac{\hat{\tau}_i}{\pi_i})E(\delta_i) = \sum_{i=1}^{N}\frac{\tau_i}{\pi_i}\pi_i = \sum_{i=1}^{N}\tau_i = \tau.$$

Note que el estimador de Horvitz-Thompson no se restringe necesariamente a un muestreo bietápico; el diseño podría ser sin problemas multietápico bajo la condición de que el total de la i-ésima UPM pueda ser insesgadamente estimado.

El siguiente teorema ilustra algunas propiedades adicionales de este estimador. Antes será necesario introducir las probabilidades conjuntas de selección de dos unidades primarias, llamadas también probabilidades de inclusión de segundo orden. Estas vienen dadas por

$$\pi_{ij} = P(\delta_i = 1, \delta_j = 1).$$

En un MASs, por ejemplo, estas probabilidades no dependen de las unidades seleccionadas y vienen dadas por $\pi_{ij} = \frac{(n-1)}{(N-1)}\frac{n}{N}$.

Proposición 4.2. *Independientemente de cómo se definan las probabilidades de inclusión de primer y segundo orden, estas deben satisfacer las siguientes propiedades:*

a)

$$\sum_{i=1}^{N} \pi_i = n$$

b)

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} \pi_{ij} = (n-1)\pi_i$$

Demostración: **a)** *Como el muestreo es sin reemplazamiento, las variables indicadoras* δ_i *satisfacen por definición*

$$\sum_{i=1}^{N} \delta_i = n.$$

Luego, tomando esperanzas

$$n = \sum_{i=1}^{N} E(\delta_i) = \sum_{i=1}^{N} \pi_i.$$

b) *Por otro lado,*

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} \pi_{ij} = \sum_{\substack{j=1 \\ j\neq i}}^{N} E(\delta_i\delta_j) = E(\delta_i(\sum_{\substack{j=1 \\ j\neq i}}^{N} \delta_j)) = E(\delta_i(n-\delta_i)) = (n-1)\pi_i. \qquad \blacksquare$$

Teorema 4.1. *Considere un muestreo bietápico que se realiza de modo que el muestreo en cualquier unidad primaria es independiente del muestreo en cualquier otra unidad primaria. Sea* $\hat{\tau}_i$ *un estimador insesgado del total* τ_i *de la* i*-ésima unidad primaria, el cual es independiente de* $\delta_1, \delta_2, \ldots, \delta_N$*. Entonces, el estimador de Horvitz-Thompson del total de la población,*

$$\hat{\tau}_{HT} = \sum_{i=1}^{N} \frac{\hat{\tau}_i}{\pi_i}\delta_i,$$

es insesgado, y su varianza viene dada por

$$V(\hat{\tau}_{HT}) = \sum_{i=1}^{N}(1-\pi_i)\frac{\tau_i^2}{\pi_i} + \sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j} + \sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i}$$

$$= \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j-\pi_{ij})(\frac{\tau_i}{\pi_i}-\frac{\tau_j}{\pi_j})^2 + \sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i}. \tag{4.6}$$

Mas aún, dos estimadores insesgados de esta varianza viene dados por

$$\hat{V}_{HT}(\hat{\tau}_{HT}) = \sum_{i=1}^{N}\frac{(1-\pi_i)}{\pi_i^2}\hat{\tau}_i{}^2\delta_i + \sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}})\frac{\hat{\tau}_i}{\pi_i}\frac{\hat{\tau}_j}{\pi_j}\delta_i\delta_j + \sum_{i=1}^{N}\frac{\hat{V}(\hat{\tau}_i)}{\pi_i}\delta_i$$

y

$$\hat{V}_{SYG}(\hat{\tau}_{HT}) = \sum_{i=1}^{N}\sum_{j>i}^{N}(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}})(\frac{\hat{\tau}_i}{\pi_i}-\frac{\hat{\tau}_j}{\pi_j})^2\delta_i\delta_j + \sum_{i=1}^{N}\frac{\hat{V}(\hat{\tau}_i)}{\pi_i}\delta_i,$$

donde $\hat{V}(\hat{\tau}_i)$ es un estimador insesgado de $V(\hat{\tau}_i)$.

<u>Demostración:</u> Puesto que estamos asumiendo esquemas sin reemplazamiento para las distintas etapas de selección, los δ_i son variables aleatorias de Bernoulli con media π_i y varianza $\pi_i(1-\pi_i)$. Estas tienen para $i \neq j$ una covarianza igual a $Cov(\delta_i,\delta_j) = \pi_{ij}-\pi_i\pi_j$. Para encontrar la varianza del estimador de Horvitz-Thompson utilizaremos la proposición 1.4, condicionando esta a la selección de las unidades primarias. Más precisamente:

$$V(\hat{\tau}_{HT}) = V(E(\hat{\tau}_{HT} \mid \delta_1,\delta_2,\ldots,\delta_N)) + E(V(\hat{\tau}_{HT} \mid \delta_1,\delta_2,\ldots,\delta_N))$$

$$= V(\sum_{i=1}^{N}\frac{E(\hat{\tau}_i)}{\pi_i}\delta_i) + E(\sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i^2}\delta_i^2)$$

$$= V(\sum_{i=1}^{N}\frac{\tau_i}{\pi_i}\delta_i) + \sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i^2}E(\delta_i^2)$$

$$= \sum_{i=1}^{N}(\frac{\tau_i}{\pi_i})^2V(\delta_i) + \sum_{\substack{i=1\\ j\neq i}}^{N}\sum_{j=1}^{N}\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}Cov(\delta_i,\delta_j) + \sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i^2}\pi_i$$

$$= \sum_{i=1}^{N}(1-\pi_i)\frac{\tau_i^2}{\pi_i} + \sum_{i=1}^{N}\sum_{\substack{i=1\\ j\neq i}}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j} + \sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i}.$$

Veamos ahora la equivalencia en (4.6) partiendo del segundo término sin el último factor $\sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i}$ que se mantiene constante en ambas formulaciones. Este término es igual a

$$\frac{1}{2}\sum_{\substack{i=1\\ j\neq i}}^{N}\sum_{j=1}^{N}(\pi_i\pi_j-\pi_{ij})(\frac{\tau_i^2}{\pi_i^2}+\frac{\tau_j^2}{\pi_j^2}-2\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}) =$$

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\pi_i\pi_j(\frac{\tau_i^2}{\pi_i^2}+\frac{\tau_j^2}{\pi_j^2})-\frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\pi_{ij}(\frac{\tau_i^2}{\pi_i^2}+\frac{\tau_j^2}{\pi_j^2})-\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}(\pi_i\pi_j-\pi_{ij})\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}$$

$$=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\pi_i\pi_j(\frac{\tau_i^2}{\pi_i^2}+\frac{\tau_j^2}{\pi_j^2})-\sum_{i=1}^{N}\tau_i^2-\sum_{i=1}^{N}\frac{\tau_i^2}{\pi_i^2}(\sum_{\substack{j=1\\ j\neq i}}^{N}\pi_{ij})-\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}(\pi_i\pi_j-\pi_{ij})\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}$$

$$=\sum_{i=1}^{N}(n-\pi_i)\frac{\tau_i^2}{\pi_i}-(n-1)\sum_{i=1}^{N}\frac{\tau_i^2}{\pi_i}-\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}(\pi_i\pi_j-\pi_{ij})\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}$$

$$=\sum_{i=1}^{N}(1-\pi_i)\frac{\tau_i^2}{\pi_i}+\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}.$$

Mostremos, finalmente, el insesgamiento de $\hat{V}_{HT}(\hat{\tau}_{HT})$. El del otro estimador queda como ejercicio. Utilizando nuevamente la proposición 1.4, se tiene que

$$E(\hat{V}_{HT}(\hat{\tau}_{HT}))=E(E(\hat{V}_{HT}(\hat{\tau}_{HT})\mid\delta_1,\delta_2,\ldots,\delta_N))$$

$$=E\left(\sum_{i=1}^{N}\frac{(1-\pi_i)}{\pi_i^2}E(\hat{\tau}_i)^2\delta_i+\sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}})E(\frac{\hat{\tau}_i}{\pi_i}\frac{\hat{\tau}_j}{\pi_j})\delta_i\delta_j+\sum_{i=1}^{N}\frac{E(\hat{V}(\hat{\tau}_i))}{\pi_i}\delta_i\right)$$

$$=E\left(\sum_{i=1}^{N}\frac{(1-\pi_i)}{\pi_i^2}(V(\hat{\tau}_i)+\tau_i^2)\delta_i+\sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}})\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}\delta_i\delta_j+\sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i}\delta_i\right)$$

$$=\sum_{i=1}^{N}\frac{(1-\pi_i)}{\pi_i^2}\tau_i^2E(\delta_i)+\sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}})\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}E(\delta_i\delta_j)+\sum_{i=1}^{N}(\frac{(1-\pi_i)}{\pi_i^2}+\frac{1}{\pi_i})V(\hat{\tau}_i)E(\delta_i)$$

$$=\sum_{i=1}^{N}(1-\pi_i)\frac{\tau_i^2}{\pi_i}+\sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}+\sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i}=V(\hat{\tau}_{HT}).\qquad\blacksquare$$

Note que aun cuando las dos formas dadas para la varianza del estimador de Horvitz-Thompson son algebraicamente las mismas y sus estimadores se basan en estas, la sustitución de las cantidades muestrales en estas expresiones sobre diseños con probabilidades no iguales proporcionan en general diferentes estimadores de esta varianza. A la segunda de estas formas se le conoce como el estimador de Sen-Yates-Grundy (SYG) y, en general, esta suele mostrar

mayor estabilidad que la primera, la cual se llama también el estimador de Horvitz-Thompson de la varianza del estimador del mismo nombre.

El estimador de Horvitz-Thompson resume prácticamente todos los estimadores de totales en esquemas sin reemplazamiento anteriormente vistos (y los de medias, al dividirlos entre la cantidad total de unidades últimas de muestreo). Un ejemplo que desarrollaremos aquí es el de un MASs. En este caso,

$$\hat{\tau}_{HT} = \sum_{i=1}^{N} \frac{\hat{\tau}_i}{\pi_i}\delta_i = \sum_{i=1}^{N} \frac{y_i}{\frac{n}{N}}\delta_i = N\bar{Y}$$

es el clásico estimador del total, cuya varianza viene dada por

$$V(\hat{\tau}) = \sum_{i=1}^{N} \frac{(1-\frac{n}{N})}{\frac{n}{N}} y_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} (\frac{\frac{n-1}{N-1} - \frac{n}{N}}{\frac{n}{N}}) y_i y_j = N^2(1-\frac{n}{N})\frac{\sigma^2}{n},$$

donde:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2 \quad \text{y} \quad \mu = \frac{1}{N}\sum_{i=1}^{n} y_i.$$

Otro caso particular, como se pide mostrar en el ejercicio 4.3 y que justifica (4.2), es el estimador para la media en un muestreo por conglomerados bietápico. En caso de estimarse el total, esta varianza viene dada por

$$V(\hat{\tau}_{HT}) = (1-\frac{n}{N})\frac{N^2}{n}\sigma_c^2 + \frac{N}{n}\sum_{i=1}^{N}(1-\frac{m_i}{M_i})M_i^2\frac{\sigma_i^2}{m_i},$$

donde $\sigma_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\tau_i - \mu_\tau)^2$.

Un problema, particular que se presenta con los estimadores de la varianza del estimador de Horvitz-Thompson es que, para algunos diseños con probabilidades distintas, estas pueden resultar negativas. A veces, la estabilidad se puede mejorar mediante una elección cuidadosa del diseño; pero en general los cálculos son complicados. Una alternativa, que evita algo de la inestabilidad potencial y la complejidad de los cálculos para la obtención de las probabilidades de inclusión, es emplear el estimador de la varianza del estimador del total considerando reemplazamiento. Esto fue lo que exactamente propusieron Hansen y Hurwitz (1943) dando pie al siguiente estimador del total que lleva sus nombres:

$$\hat{\tau}_\psi = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}}{\psi_i},$$

donde ψ_i es la probabilidad de tomar la unidad primaria i en una selección (no interesa cuál); δ_i es el número de veces que la unidad i es seleccionada en la muestra, y los $\hat{\tau}_{ij}$ son

estimadores insesgados del total de la unidad primaria i, para la j-ésima selección de dicha unidad. Note que el estimador de Horvitz-Thompson resulta de esta expresión si sustituimos arriba a ψ_i por un promedio de elegir la unidad i en una extracción; vale decir, por $\frac{\pi_i}{n}$. Se puede probar (véase el ejercicio 4.12) que un estimador insesgado de la varianza de $\hat{\tau}_\psi$ viene dado por

$$\hat{V}(\hat{\tau}_\psi) = \frac{1}{n(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{\delta_i} (\frac{\hat{\tau}_{ij}}{\psi_i} - \hat{\tau}_\psi)^2. \tag{4.7}$$

4.10. Muestreo ppt

Un caso particular de muestreo con probabilidades desiguales es el del muestreo con probabilidades proporcionales al tamaño (ppt). Si X_i denota el tamaño (valor de alguna variable cuantitativa) de una unidad i, entonces la probabilidad de que se seleccione esta unidad en el muestreo ppt será proporcional a X_i, digamos $\pi_i^0 = CX_i$, donde C es una constante de proporcionalidad. Dado que por la proposición 4.2 $\sum_{i=1}^{N} \pi_i^0 = n$, resulta que de reemplazarse las probabilidades anteriores en esta igualdad uno obtiene que $C = \frac{n}{\sum_{i=1}^{N} X_i}$ y, por tanto:

$$\pi_i^0 = \frac{X_i}{\sum_{j=1}^{N} X_j} n.$$

Esto, sin embargo, podría generar una cantidad mayor que 1, si la unidad i es relativamente grande. En tal caso, las probabilidades se fijan en 1 (y, consecuentemente las unidades correspondientes serán siempre seleccionadas); mientras que las probabilidades de las demás unidades se deben reescalar para que queden bien definidas; más explícitamente, si al conjunto de K unidades en el subconjunto $\mathcal{K}$ de $\mathcal{P}$ les corresponde por lo anterior una probabilidad de 1, entonces cualquier unidad $i \notin \mathcal{K}$ tendrá una probabilidad de selección igual a

$$\frac{X_i(n-K)}{\sum_{j \notin \mathcal{K}} X_j}.$$

Así, las probabilidades de inclusión de primer orden en un muestreo ppt quedan, finalmente, definidas por:

$$\pi_i = \begin{cases} 1 & \text{si } \pi_i^0 \geq 1 \\ \frac{X_i(n-K)}{\sum_{j \notin \mathcal{K}} X_j} & \text{si } \pi_i^0 < 1 \end{cases}$$

Como se aprecia, las probabilidades de inclusión de primer orden son sencillas de obtener. En R estas se calculan con el paquete sampling de R bajo el comando `inclusionprobabilities`. Para ilustrar el cálculo, adelantémos un poco al ejemplo 4.6, en el cual se nos pide seleccionar una muestra ppt de tamaño 3 basándonos en el tamaño del terreno en m^2 que ocupan 6 supermercados de un consorcio en una ciudad. El código es

```
X = c(300,200,100,1000,150,500)
pik = inclusionprobabilities(X,3)
pik

## [1] 0.48 0.32 0.16 1.00 0.24 0.80
```

Note que bajo este esquema el supermercado D resultará ser siempre seleccionado.

A diferencia de las probabilidades de inclusión de primer orden, las de segundo, que son indispensables por el teorema 4.1 en la obtención de las estimaciones de la varianza del estimador, no solo no son únicas sino que son difíciles de obtener. Estas probabilidades deben satisfacer la proposición 4.2 b), lo cual nos conduce en general a resolver sistemas de ecuaciones nada triviales. En la práctica, la obtención de estas probabilidades es todo un desafío; por ello que en lugar de buscar fijarlas y con estas estimar la varianza del estimador de interés (sin que esto nos diga cómo obtener la muestra), es mucho más conveniente tratar de prescindir de estas, ya sea tomando un muestreo con reemplazamiento o diseñando esquemas de muestreo sin reemplazamiento que respeten las probabilidades de inclusión de primer orden y que satisfagan 4.2 b). Algunos de estos esquemas se tratarán en la sección 4.12.

4.11. Muestreo secuencial ppt

Otra alternativa cercana al muestreo ppt es el muestreo secuencial ppt. Si las probabilidades de selección fuesen iguales, sabemos que seleccionar simultáneamente al azar y sin reemplazamiento n unidades de una población equivale a seleccionar secuencialmente una por una y sin reemplazamiento cada una de las n unidades. Esto es falso en un muestreo con probabilidades desiguales. El muestreo secuencial ppt puede verse como el esquema secuencial último, donde la probabilidad de que se selecciones la unidad i en la primera ocasión es $\frac{X_i}{\sum_{j=1}^{N} X_j}$. Sin embargo, como el muestreo es sin reemplazamiento, la probabilidad de que se seleccione la unidad j para la segunda ocasión dependerá de la unidad i seleccionada en la primera. Formalmente, si denotamos por $\pi_i(m)$ la probabilidad de que la unidad i sea seleccionada en una muestra secuencial ppt de tamaño m y $X = \sum_{j=1}^{N} X_j$, entonces

$$\pi_i(1) = \frac{X_i}{X}$$

y

$$\pi_i(m) = \pi_i(m-1) + P(\cap_{\ell=1}^{m-1} E_{\ell,i}^c \cap E_{m,i}) = \pi_i(m-1) + P(E_{m,i} \mid \cap_{\ell=1}^{m-1} E_{\ell,i}^c) P(\cap_{\ell=1}^{m-1} E_{\ell,i}^c)$$

$$= \pi_i(m-1) + \sum_{i_1,i_2,\ldots,i_{m-1}} \left(\prod_{\ell=1}^{m-1} \frac{X_{i_\ell}}{X - \sum_{h=1}^{\ell-1} X_{i_h}} \right) \frac{X_i}{X - \sum_{h=1}^{m-1} X_{i_h}}, \tag{4.8}$$

donde $E_{k,i}$ denota el evento en que la unidad i es seleccionada en la k-ésima selección, la sumatoria de orden $m-1$ recorre todos los índices de 1 hasta N, sin incluir al término i y sin repeticiones, y la sumatoria desde 1 hasta 0 se conviene que es 0.

En los casos particulares de $m=1$, $m=2$ y $m=3$, estos desarrollos vienen dados por

$$\pi_i(1) = \frac{X_i}{X}$$

$$\pi_i(2) = \pi_i(1) + \sum_{\substack{i_1=1 \\ i_1 \neq i}}^{N} \left(\frac{X_i}{X - X_{i_1}}\right)\left(\frac{X_{i_1}}{X}\right)$$

$$\pi_i(3) = \pi_i(2) + \sum_{\substack{i_1=1 \\ i_1 \neq i_2 \neq i}}^{N} \sum_{i_2=1}^{N} \left(\frac{X_i}{X - X_{i_1} - X_{i_2}}\right)\left(\frac{X_{i_2}}{X - X_{i_1}}\right)\left(\frac{X_{i_1}}{X}\right).$$

Note que en el caso especial en que las X_i sean todas iguales, uno obtiene un MASs. Aquí la fórmula (4.8) se reduce a $\pi_i(m) = \frac{m}{N}$, cualquiera sea el valor de $i = 1, 2, \ldots, N$ y del tamaño de muestra m.

Ejemplo 4.5. *Un grupo comercial posee 6 supermercados en una ciudad, los cuales ocupan terrenos con tamaños de entre 100 y 1000 metros cuadrados. Se desea estimar la cantidad total de ventas mensual para el grupo en la ciudad, para lo cual se seleccionarán al azar y sin reemplazamiento tres de estos supermercados. Si, para fines didácticos, dispusiéramos de la siguiente información:*

Supermercado	*Tamaño (m^2)*	*Ventas totales en miles de dólares*
A	*300*	*24*
B	*200*	*20*
C	*100*	*11*
D	*1000*	*245*
E	*150*	*18*
F	*500*	*90*

Obtenga, usando un muestreo secuencial ppt y un muestreo ppt, una estimación del total buscado junto con una estimación de su error estándar de estimación. Replique todo lo anterior si el muestreo fuese con reemplazamiento.

Solución: Notemos que el verdadero total de ventas a estimar para la cadena es de 408 000 dólares. Consideremos primero el muestreo secuencial ppt, para el cual hemos desarrollado la siguiente función en R que calcula sus probabilidades de inclusión de primer y segundo orden.

```
library(combinat) # Requiere del paquete combinat
pisppt <-function(X,n){
N = length(X)
XT = sum(X)
m = apply(combn(X,n),2,permn)
m = matrix(unlist(m),ncol=n,byrow=TRUE)
nm = dim(m)[1] # Numero de permutaciones de N en n
p<-pi1<-0
for (j in 1:nm){
  p[j] = prod(m[j,])/(XT*prod(XT-cumsum(m[j,1:n-1])))}
pi2=matrix(0,N,N)
for (i in 1:(N-1)){
   aux1 = (m==X[i])
   index = which(apply(1*aux1,1,sum)==1)
   pi1[i] = sum(p[index])
   for (j in (i+1):N){
      aux2 = (m==X[j])
      aux2 = 1*aux2[index,]
      pi2[i,j] = sum(p[index[which(apply(aux2,1,sum)==1)]])}}
pi1[N] = n-sum(pi1)
pi2 = pi2+t(pi2)
diag(pi2) = pi1
pi2}
```

Una aplicación de esta función nos brinda las siguientes probabilidades de inclusión de primer y segundo orden, donde las primeras se encuentran en la diagonal de la matriz.

```
(p = pisppt(X,3))

##        [,1]   [,2]   [,3]  [,4]   [,5]  [,6]
## [1,] 0.5234 0.1223 0.0602 0.451 0.0910 0.323
## [2,] 0.1223 0.3743 0.0392 0.315 0.0594 0.213
## [3,] 0.0602 0.0392 0.1982 0.162 0.0291 0.106
## [4,] 0.4507 0.3146 0.1624 0.899 0.2401 0.631
## [5,] 0.0910 0.0594 0.0291 0.240 0.2893 0.159
## [6,] 0.3228 0.2132 0.1056 0.631 0.1591 0.716
```

Note, por ejemplo, que la probabilidad de que se seleccione al supermercado D es bastante alta e igual a 0.899; mientras que la probabilidad de que este supermercado sea seleccionado en la muestra de tres supermercados junto con, digamos, el supermercado A es de 0.4507.

Para la selección de la muestra debemos proceder secuencialmente y tomar tres números aleatorios. Supongamos que nos salieron 0.8869, 0.9493 y 0.4259. El primer supermercado seleccionado será por la tabla

Supermercado	*Tamaño*	$\pi_i(1)$	$\Pi_i(1)$
A	*300*	*0.133333333*	*0.133333333*
B	*200*	*0,088888889*	*0.222222222*
C	*100*	*0.044444444*	*0.266666667*
D	*1000*	*0.444444444*	*0.711111111*
E	*150*	*0.066666667*	*0.777777778*
F	*500*	*0.222222222*	*1*
Total	*2250*		

el supermercado F. Cabe aclarar que $\Pi_i(1)$ *denota aquí la probabilidad acumulada para la primera de selección. Eliminado el supermercado F del proceso, el segundo supermercado seleccionado será por la tabla*

Supermercado	*Tamaño*	$\pi_{i\|6}(2)$	$\Pi_{i\|6}(2)$
A	*300*	*0.171428571*	*0.171428571*
B	*200*	*0.114285714*	*0.285714286*
C	*100*	*0.057142857*	*0.342857143*
D	*1000*	*0.571428571*	*0.914285714*
E	*150*	*0.085714286*	*1*
Total	*1750*		

el supermercado E. Finalmente, eliminados los dos supermercados ya seleccionados, el último supermercado seleccionado será por la tabla

Supermercado	*Tamaño*	$\pi_{i\|5,6}(3)$	$\Pi_{i\|5,6}(3)$
A	*300*	*0.1875*	*0.1875*
B	*200*	*0.125*	*0.3125*
C	*100*	*0.0625*	*0.375*
D	*1000*	*0.625*	*1*
Total	*1600*		

el supermercado D. Con ellos, la estimación pedida será de

$$\hat{\tau} = \frac{90}{0.7155999} + \frac{18}{0.2893101} + \frac{245}{0.8991226} = 460.47345$$

en miles de dólares. Este mismo resultado puede obtenerse de manera más directa con R y el paquete sampling a través de los códigos

```
y = c(24,20,11,245,18,90)
HTestimator(y[4:6],diag(p)[4:6])

##      [,1]
## [1,]  460
```

Más aún, el error estándar de estimación estimado de esta estimación puede obtenerse del teorema 4.1 con la función varHT mediante

```
pik2 = p[4:6,4:6]
sqrt(varHT(y[4:6],pik2,1))

## [1] 76.1

sqrt(varHT(y[4:6],pik2,2))

## [1] 73.1
```

donde el primer término corresponde a la estimación con el estimador de Horvitz-Thompson; mientras que el segundo está asociado al método SGY.

Enfoquémonos ahora en el muestreo ppt, para el que ya obtuvimos las probabilidades de inclusión de primer orden y en el que el supermercado D sería de todas maneras seleccionado. La dificultad consiste en hallar las probabilidades de inclusión de segundo orden. Según la proposición 4.2 b), ellas deben satisfacer las siguientes ecuaciones:

$$\begin{aligned}
\pi_{12}+\pi_{13}+\pi_{14}+\pi_{15}+\pi_{16} &= 0.96\\
\pi_{21}+\pi_{23}+\pi_{24}+\pi_{25}+\pi_{26} &= 0.64\\
\pi_{31}+\pi_{32}+\pi_{34}+\pi_{35}+\pi_{36} &= 0.32\\
\pi_{41}+\pi_{42}+\pi_{43}+\pi_{45}+\pi_{46} &= 2\\
\pi_{51}+\pi_{52}+\pi_{53}+\pi_{54}+\pi_{56} &= 0.48\\
\pi_{61}+\pi_{62}+\pi_{63}+\pi_{64}+\pi_{65} &= 1.6
\end{aligned}$$

Este sistema posee, sin embargo, infinitas soluciones, una de las cuales se resume en la siguiente matriz $\boldsymbol{\pi} = [\pi_{ij}]$ *de probabilidades de inclusión de segundo orden:*

$$\boldsymbol{\pi} = \begin{bmatrix}
0 & 0.05 & 0.04 & 0.48 & 0.04 & 0.35\\
0.05 & 0 & 0.02 & 0.32 & 0.03 & 0.22\\
0.04 & 0.02 & 0 & 0.16 & 0.02 & 0.08\\
0.48 & 0.32 & 0.16 & 0 & 0.24 & 0.8\\
0.04 & 0.03 & 0.02 & 0.24 & 0 & 0.15\\
0.35 & 0.22 & 0.08 & 0.8 & 0.15 & 0
\end{bmatrix}$$

Esta elección arbitraria debería corresponder al mecanismo de selección que se empleará para obtener simultáneamente las tres unidades pedidas (en la que el supermercado D estará de todos modos). Tal mecanismo es difícil de deducir, por lo cual una mejor estrategia sería la de primero fijar el mecanismo de selección para luego encontrar la matriz particular $\boldsymbol{\pi}$ *asociada. Ello es lo que precisamente haremos en la siguiente sección.*

Analicemos, finalmente, la posibilidad de tomar un muestreo con reemplazamiento, el cual simplifica muchísimo el proceso de selección. Supongamos para ello que obtuvimos los números aleatorios 0.09245, 0.7779 y 0.5865. Entonces, de la primera tabla obtenida en el muestreo secuencial ppt (con $\psi_i = \pi_i^1$*), los supermercados seleccionados serán A, D y F. Ello nos da una estimación para el total de ventas de*

$$\hat{\tau_\psi} = (\frac{24}{0.133} + \frac{245}{0.444} + \frac{90}{0.222})/3 = 378.75$$

en miles de dólares. El error estándar de estimación estimado a través de la fórmula (4.7) es de 194.556 en miles de dólares. □

4.12. Muestreo sin reemplazamiento con probabilidades desiguales

El ejemplo anterior ilustra la complejidad no solo del cálculo de las probabilidades de selección, sino de la selección misma de la muestra. Como adelantamos, veremos en esta sección esquemas de muestreo alternativos, los cuales podrían suplir la metodología anteriormente descrita. Estos esquemas buscan respetar las probabilidades de inclusión pre-definidas π_i, sobre una muestra de tamaño n, y en algunos casos nos brindan probabilidades de inclusión de segundo orden. Para formalizarlas, definamos un diseño de muestreo como el par $(\mathcal{Q}, p)$, donde $\mathcal{Q}$ denota el conjunto de todas las muestras posibles de tamaño n que se pudieran extraer de la población $\mathcal{P} = \{1, 2, \ldots, N\}$ y p denota a una distribución de probabilidades (conjunta) sobre las muestras en $\mathcal{Q}$; esto es,

$$0 < p(\boldsymbol{\delta}) \leq 1, \forall \boldsymbol{\delta} \in \mathcal{Q} \quad \text{y} \quad \sum_{\boldsymbol{\delta} \in \mathcal{Q}} p(\boldsymbol{\delta}) = 1.$$

En el muestreo con probabilidades iguales, por ejemplo, los diseños de muestreo más empleados son los correspondientes al MASc, caracterizado por $\mathcal{Q} = \mathcal{R} = \{\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_N) \in \mathcal{P}^N / \sum_{i=1}^N \delta_i = n\}$, y al MASs, caracterizado por $\mathcal{Q} = \mathcal{S} = \{\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_N) \in \{0,1\}^N / \sum_{i=1}^N \delta_i = n\}$. La cardinalidad de estos conjuntos viene dada, respectivamente, por $\#(\mathcal{R}) = C_n^{N+n-1}$ y $\#(\mathcal{S}) = C_n^N$; mientras que sus distribucions de probabilidade son iguales a $p(\boldsymbol{\delta}) = \frac{1}{C_n^{N+n-1}}$ y $p(\boldsymbol{\delta}) = \frac{1}{C_n^N}$, respectivamente.

La definición de un diseño de muestreo no nos dice, sin embargo, cómo es que en la práctica uno podría tomar la muestra. Para ello requeriremos de lo que se conoce como un algoritmo

de muestreo (*sampling algorithm*); esto es, un procedimiento que nos permita seleccionar la muestra. La manera más directa de definir este es la enumerativa. Esta consiste en listar todos los elementos del diseño $(\mathcal{Q}, p)$ junto con sus probabilidades acumuladas y luego seleccionar aquel elemento en $\mathcal{Q}$ cuya probabilidad acumulada sea la primera en superar a cierto número aleatorio. Tal algoritmo resulta, sin embargo, prohibitivo si n y N son relativamente grandes o si p no está completamente especificada. El texto de Tillé (2006) se enfoca precisamente en el desarrollo de algoritmos de muestreo que sean más eficientes que el enumerativo planteado. Algunos de estos esquemas se introducen seguidamente.

4.12.1. El esquema de Poisson

Este es uno de los esquemas más simples, pero tiene la desventaja de proveer un tamaño de muestra aleatorio y no fijo. El algoritmo consiste en generar N números aleatorios $u_1, u_2, \ldots, u_N$ uniformes en el intervalo unitario y seleccionar en la muestra a la unidad i si $u_i < \pi_i$. Así, si δ_i denota la variable binaria que indica si la unidad i es seleccionada o no, el tamaño de muestra bajo este esquema será $n_s = \sum_{i=1}^{N} \delta_i$ y su número esperado $E(n_s) = \sum_{i=1}^{N} \pi_i$. En el caso particular de un muestreo ppt con $\pi_i = \frac{X_i n}{X}$, el tamaño de muestra esperado será precisamente el planificado n. Este esquema se encuentra implementado en el paquete sampling de R bajo el comando `UPpoisson,` donde el prefijo UP significa "unequal probabilities".

4.12.2. El esquema sistemático ppt

Este es uno de los esquemas de muestreo ppt más populares y simples para un tamaño de muestra fijo n. Su lógica es la misma que la del muestreo sistemático; esto es, en él se selecciona al azar un único número aleatorio u en el intervalo unitario y a partir de este se hacen sistemáticamente las siguientes selecciones. Si definimos por $\Pi_i = \sum_{k=1}^{i} \pi_k$ a la suma acumulada de las probabilidades de selección, la primera unidad j_1 que se seleccionará será la primera en $\mathcal{P}$ para la cual se cumpla que $0 \leq u < \Pi_{j_1}$. De manera similar la k-ésima unidad a seleccionarse, j_k, será aquella que satisfaga $\Pi_{j_k - 1} \leq u + k - 1 < \Pi_{j_k}$, donde $k = 2, \ldots, n$. Se puede demostrar que las probabilidades de inclusión de segundo orden en este esquema vienen para $i < j$ dadas por

$$\pi_{ij} = \text{mín}\{\text{máx}\{0, \pi_i - D_{ij}\}, \pi_j\} + \text{mín}\{\pi_i, \text{máx}\{0, D_{ij} + \pi_j - 1\}\},$$

donde $D_{ij} = V_{ij} - [V_{ij}]$, siendo $[.]$ la notación para el máximo valor entero y $V_{ij} = \sum_{k=i}^{j-1} \pi_k$.

Una desventaja de este esquema es que muchas de las probabilidades anteriores son nulas. Para atenuar ello y hacer que el esquema no sea dependiente del orden dado en el marco muestral, uno podrá aplicarlo luego de ordenar aleatoriamente el marco muestral, de tal manera que las probabilidades de inclusión de segundo orden sean las medias de las

probabilidades de inclusión del esquema sistemático anterior para todas las permutaciones posibles en el marco muestral. Claramente, esto será posible si el aspecto computacional lo permite; es decir, si el tamaño de la población no es muy grande. El esquema anterior así como este último, se encuentran implementados en el paquete sampling de R a través de los comandos `UPsystematic` y `UPrandomsystematic`, respectivamente. Se dispone también del comando `UPsystematicpi2` que calcula, para el primero, las probabilidades de inclusión de segundo orden.

4.12.3. El esquema de Sampford

Este es un diseño sin reemplazamiento que destaca por su simplicidad y, como Sampford (1967) lo deriva, nos provee de probabilidades de inclusión de segundo orden explícitas. Este es un esquema de rechazo que consiste en seleccionar la primera unidad con probabilidades $\frac{\pi_i}{n}$ y las demás $n-1$ unidades con reemplazamiento y probabilidades proporcionales a $\frac{\pi_i}{1-\pi_i}$. Esta muestra se acepta luego si las n unidades son todas distintas, y se rechaza en caso contrario; el proceso se repite hasta alcanzar las n muestras requeridas. Con este esquema se presenta un problema cuando las probabilidades de inclusión son grandes, pues de elegirse inicialmente las unidades con estas probabilidades, las demás $n-1$ difícilmente podrán ser distintas. El método se encuentra implementado en el paquete sampling de R a través de los comandos `UPsampford` y `UPsampfordpi2` ; este último nos provee de las probabilidades de inclusión de segundo orden.

4.12.4. Esquemas de división

Deville y Tillé (1998) propusieron un esquema general sin reemplazamiento para un tamaño de muestra fijo n que se basa en la idea de expresar el vector de probabilidades de inclusión de primer orden $\boldsymbol{\pi}$ como una combinación lineal convexa de M vectores similares $\boldsymbol{\pi}^{(1)}(0), \boldsymbol{\pi}^{(2)}(0), \ldots, \boldsymbol{\pi}^{(M)}(0)$ bajo escalares $\lambda_1(0), \lambda_2(0), \ldots, \lambda_M(0) \in [0,1]$:

$$\boldsymbol{\pi} = \sum_{j=1}^{M} \lambda_j(0)\boldsymbol{\pi}^{(j)}(0),$$

de tal manera que este vector se actualize para el paso 1 como uno de los M vectores anteriores, digamos $\boldsymbol{\pi}(1) = \boldsymbol{\pi}^{(k)}(0)$, el cual será seleccionado con probabilidad $\lambda_k(0)$. El vector resultante tomará ahora el rol del vector de probabilidades de inclusión de primer orden y el algoritmo se repetirá hasta el paso K en el que $\boldsymbol{\pi}(K) \in \{0,1\}^N$, lo cual nos brindará la muestra buscada. Dependiendo de cómo se especifique la combinación lineal convexa en cada paso t

$$\boldsymbol{\pi}(t) = \sum_{j=1}^{M} \lambda_j(t)\boldsymbol{\pi}^{(j)}(t), \tag{4.9}$$

el método generará una gran variedad de esquemas distintos. Aquí solo explicitaremos algunos de ellos

El esquema por división hacia un MAS

Este esquema considera $M = 2$ y fuerza a que uno de los dos vectores de mezcla en (4.9) corresponda siempre a un MAS. El escalar $\lambda_1(t)$ se escoge de tal manera que, en la siguiente iteración, la probabilidad de inclusión de la unidad k con el valor más cercano a 0 o 1 tome precisamente uno de estos valores. Dado que toda unidad con una probabilidad de inclusión de 0 o 1 no integrará o integrará con certeza la muestra final, el algoritmo se simplifica para cada iteración.

El esquema pivotal

Este esquema considera $M = 2$ y tiene la peculiaridad de que modifica en cada paso solamente las probabilidades de inclusión de dos de sus unidades. Si en el paso t se eligen las unidades i y j (de probabilidades no nulas ni 1), entonces dependiendo si $\pi_i(t) + \pi_j(t)$ es estrictamentes mayor que 1 o no, el esquema se define por

$$\lambda(t) = \frac{1 - \pi_j(t)}{2 - \pi_i(t) - \pi_j(t)},$$

$$\pi_k^{(1)}(t) = \begin{cases} \pi_k(t) & \text{si } k \in \mathcal{P} \setminus \{i, j\} \\ 1 & \text{si } k = i \\ \pi_i(t) + \pi_j(t) - 1 & \text{si } k = j \end{cases}$$

y

$$\pi_k^{(2)}(t) = \begin{cases} \pi_k(t) & \text{si } k \in \mathcal{P} \setminus \{i, j\} \\ \pi_i(t) + \pi_j(t) - 1 & \text{si } k = i \\ 1 & \text{si } k = j \end{cases}$$

o, respectivamente, por

$$\lambda(t) = \frac{\pi_i(t)}{\pi_i(t) + \pi_j(t)},$$

$$\pi_k^{(1)}(t) = \begin{cases} \pi_k(t) & \text{si } k \in \mathcal{P} \setminus \{i, j\} \\ \pi_i(t) + \pi_j(t) & \text{si } k = i \\ 0 & \text{si } k = j. \end{cases}$$

y

$$\pi_k^{(2)}(t) = \begin{cases} \pi_k(t) & \text{si } k \in \mathcal{P} \setminus \{i, j\} \\ 0 & \text{si } k = i \\ \pi_i(t) + \pi_j(t) & \text{si } k = j. \end{cases}$$

En el primer caso se fija una probabilidad de 1 a una sola de las unidades; mientras que en el segundo se fija una probabilidad de 0 a solo una de las unidades. De esta manera, el esquema requiere de a lo más N pasos para obtener la muestra.

Tanto el esquema pivotal como otros de división en M clases, como el **esquema de eliminación de Tillé** o el **esquema de Midzuno generalizado**, se encuentran implementados

en el paquete sampling de R. Mayores detalles de estos y otros esquemas se pueden encontrar en el texto de Tillé (2006).

Ejemplo 4.6. *Estimemos, bajo los esquemas ppt dados, el total de ventas para los supermercados del ejemplo 4.5. La estimación de Horvitz-Thompson se obtiene mediante*

```
y = c(24,20,11,245,18,90)
m = UPpoisson(pik)
HTPoisson = HTestimator(y[m==1],pik[m==1])
m = UPsystematic(pik)
HTsys = HTestimator(y[m==1],pik[m==1])
m = UPrandomsystematic(pik)
HTrsys = HTestimator(y[m==1],pik[m==1])
m = UPsampford(pik)
HTsam = HTestimator(y[m==1],pik[m==1])
c(HTPoisson,HTsys,HTrsys,HTsam)

## [1] 358 420 432 389
```

□

4.13. Muestreo por conglomerados para la población api

Para ilustrar el uso del paquete survey en el muestreo por conglomerados consideremos nuevamente la base de datos poblacional api y tomaremos como conglomerados a los distritos escolares (variable dnum).

```
library(survey)
data(api)
K = dim(apipop)[1]
apipop$dnum[1:100] # mostrando parte de la variable de conglomeración

##   [1]   6   6   6   6   6   6   6   6   6   6   6   6   6   6   6   6   7
##  [18]   7   7   7  60  60  60  60  60  60  60  60  60  60  60  60  60  60
##  [35] 116 116 116 116 116 116 116 116 116 116 116 211 211 211 248 248 248
##  [52] 248 248 248 248 248 248 248 248 248 248 248 248 248 248 248 248 248
##  [69] 248 248 248 248 248 248 248 248 248 248 248 248 248 248 248 248 248
##  [86] 248 294 294 294 294 294 294 294 294 294 294 294 294 294 294
```

```
(N  = length(table(apipop$dnum)))

## [1] 757
```

Como se ve, existen 757 distritos escolares o conglomerados.

Supongamos ahora que deseamos realizar un muestreo por conglomerados de una etapa mediante una selección de 15 distritos escolares. La obtención de esta muestra no es tan directa, pero por fortuna podemos invocar al paquete sampling de R. Este paquete contiene la rutina `cluster` , que permite obtener muestras por conglomerados. Los códigos del caso son

```
library(sampling)
n = 15
set.seed(12345)
aux1=cluster(apipop,clustername=c("dnum"),n, method=c("srswor"),description=T)

## Number of selected clusters: 15
## Number of units in the population and number of selected units: 6194 103

samplec1 = getdata(apipop, aux1)
L = dim(aux1)[1]
```

El diseño se completará con

```
(dclus1<-svydesign(ids=~dnum, fpc=rep(N,L), data=samplec1))

## 1 - level Cluster Sampling design
## With (15) clusters.
## svydesign(ids = ~dnum, fpc = rep(N, L), data = samplec1)
```

Note que este es un diseño sin reemplazamiento, pues se incluye un factor de corrección para poblaciones finitas. Algo que remarcar aquí y que no ocurría en los diseños anteriores es la presencia de `ids=~dnum` , que especifica a la variable `dnum` como variable de conglomeración.

Analicemos ahora, como en los diseños previos, cómo estimar el número total de matriculados y la media del índice api para el año 2000:

```
svytotal(~enroll,dclus1)

##          total      SE
## enroll 3219521 1211326
```

```
svymean(~api00,dclus1)

##       mean   SE
## api00  724 26.3
```

Note que este diseño resulta ser menos preciso que los diseños MASs y MAE vistos anteriormente.

Consideremos ahora un muestreo aleatorio por conglomerados bietápico con 40 unidades primarias (distritos escolares) y 5 unidades secundarias (colegios) por distrito. Si bien el paquete survey contiene una base de datos con estas características llamada `apiclus2`, nosotros buscaremos tomar una muestra propia. Para esto podríamos apelar al comando `mstage` del paquete sampling, que en teoría permite obtener este tipo de muestras. Tal estrategia, sin embargo, no será aquí conveniente ya que los argumentos de dicho comando incluyen al número exacto de unidades secundarias a tomar dentro de cada unidad primaria, el cual es a priori desconocido al no tener todos los distritos escolares 5 o más colegios. Nuestra muestra bietápica la obtendremos más bien con la siguiente rutina que solo hace uso del comando `cluster`

```
set.seed(12345)
Pop = apipop
aux0 = aggregate(Pop[,6],by=list(Pop$dnum),function(x)x[1])
aux1 = aggregate(Pop[,7],by=list(Pop$dnum),length)
Popd = cbind(aux0,aux1) # Se crea una nueva base de datos de distritos
names(Popd)[c(2,4)]=c("dname","Ncdis")
Pop = merge(Pop,Popd[,c(2,4)],by=c("dname"))
m1<-sampling:::cluster(Pop,clustername=c("dnum"),size =40,method ="srswor")
m1<-getdata(Pop,m1)  # Muestra en la primera etapa (distritos)
t = as.numeric(sapply(table(m1$dnum),function(x) min(5,x)))
m2 = NULL
for(i in 1:40){ # Muestra en la segunda etapa (colegios)
  mx = m1[m1$dnum==unique(m1$dnum)[i],]
  mx$Prob1 = mx$Prob
  m<-sampling:::cluster(mx,clustername=c("snum"),size=t[i],method ="srswor")
  m = getdata(mx,m)
  m2 = rbind(m2,m)}
m2$w = 1/(m2$Prob1*m2$Prob) # Pesos de muestreo
m2$fpc1 = fpc=rep(N,dim(m2)[1])
```

El objeto diseño apropiado con la metadata necesaria para este ejemplo es

```
(dclus2 <- svydesign(ids=~dnum+snum,fpc=~fpc1+Ncdis,data=m2))

## 2 - level Cluster Sampling design
## With (40, 129) clusters.
## svydesign(ids = ~dnum + snum, fpc = ~fpc1 + Ncdis, data = m2)
```

Analicemos ahora, como en los diseños previos, la estimación del número total de matriculados y la media del índice api para el 2000:

```
svytotal(~enroll, dclus2, na.rm=TRUE)

##          total     SE
## enroll 3059677 651303

svymean(~api00, dclus2)

##       mean   SE
## api00  702 20.1
```

4.14. Diseño por conglomerados ppt para la población penal

Nuestro interés en esta sección será planificar una futura encuesta por muestreo para la población penal del Perú con el fin de estimar, con la mayor precisión posible, la proporción de internos sentenciados dadas ciertas restricciones de presupuesto. Para ello propondremos un diseño por conglomerados bietápico en el que seleccionaremos las unidades primarias, que estarán constituidas por los establecimientos penales (EP), con probabilidades proporcionales a su número de internos y luego tomaremos internos mediante un MASs. Aquí consideraremos los EP definidos en el capítulo 3 y excluiremos a los penales de Barbadillo y la Base Naval del Callao.

La pregunta central es entonces cuántos EP e internos se deben seleccionar. La respuesta a ello no es trivial, ya que el muestreo ppt de primera etapa no solo nos inhibe de utilizar los resultados de la sección 4.8, sino que no nos provee de fórmulas explícitas para la varianza de la estimación de nuestra proporción buscada. Requeriremos, asimismo, de estimaciones de la proporción de sentenciados, las cuales las tomaremos del censo del 2016. Detalles de la base de datos, costos estimados y cálculo de las proporciones comentadas se muestran en el siguiente código:

```
load("cp16f.RData")
cp16x = cp16f[-which(cp16f$EP=="Barbadillo"),]
cp16x = cp16x[-which(cp16x$EP=="Base Naval Callao"),]
pa = by(cp16x$SITUACION_JURIDICA,cp16x$EP,table)
cEP1 = unlist(lapply(pa,"[[",1))
cEP2 = unlist(lapply(pa,"[[",2))
pEPs = as.vector(cEP2/(cEP1 + cEP2)) # prop. de sentenciados por EP
M = as.vector(unlist(table(droplevels(cp16x$EP)))) # num.de internos por EP
N = length(M) # numero de EP's
c2 = rep(5,N)
c2[c(8,13:18)]=3
cc = c(750,c2,10000)
library(nloptr)
```

donde en las últimas filas hemos estimado un costo por EP a seleccionar de 750 soles, un costo por interno de 5 soles (con excepción de Lima y Callao, en que este se reduce a 3 soles) y un presupuesto total para el trabajo de campo de máximo 10 000 soles.

Un aspecto clave para sugerir los tamaños de muestra será calcular una estimación de la varianza de la proporción de sentenciados a estimar. Para ello utilizaremos, por simplicidad, un esquema sistemático ppt, el cual recordemos nos provee de las probabilidades de inclusión de segundo orden que son esenciales en el cálculo del estimador de Horvitz-Thompson. Dado entonces el número de EP a seleccionar (n), la cantidad de internos por EP a tomar (m), la cantidad de internos por EP (M) y las proporciones de internos sentenciados estimadas por EP ($pEPs$), la función siguiente permite calcular la varianza (4.6) en discusión

```
Vem <-function(m,n,M,pEPs,cc){ N = length(M) # número de EP's
pik = inclusionprobabilities(M,n)
pik2 = UPsystematicpi2(pik)
K = sum(M) # número total de internos
v1 = 0;v2 = sum((1-m/M)*(M^3)*pEPs*(1-pEPs)/((M-1)*m*pik))
for(i in 1:(N-1)){
for(j in (i+1):N){
v1=v1+(pik[i]*pik[j]-pik2[i,j])*((M[i]*pEPs[i]/pik[i]-M[j]*pEPs[j]/pik[j])^2)
}}
(v1 + v2)/(K^2)}
```

Nuestro diseño buscará minimizar la varianza anterior, sujeto a que los costos de muestreo

no superen el presupuesto otorgado. Sin embargo, dado que este costo

$$\sum_{i=1}^{N} c_1\delta_i + \sum_{i=1}^{N}\sum_{j=1}^{M_i} c_{2i}\delta_{j|i}\delta_i$$

es aleatorio, consideraremos su costo esperado. Concretamente, nuestro problema se reducirá a resolver, con respecto a n y los m_i la minimización de

$$\frac{1}{K^2}\left(\sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij})(\frac{M_i\bar{p}_i}{\pi_i} - \frac{M_j\bar{p}_j}{\pi_j})^2 + \sum_{i=1}^{N} M_i^3(1-\frac{m_i}{M_i})\frac{\bar{p}_i(1-\bar{p}_i)}{m_i(M_i-1)\pi_i}\right), \tag{4.10}$$

sujeto a que $c_1 n + \sum_{i=1}^{N} c_{2i}m_i\pi_i \leq C_0$, $m_i \leq M_i$ y $n \leq N$. Aquí, c_1 denota el costo por EP seleccionado, c_{2i} el costo unitario por interno dentro del EP i y C_0 el presupuesto total para el trabajo de campo. Note que el problema (4.10) es uno de programación no lineal entera con restricciones de desigualdad. Aquí, los π_i y π_{ij} dependen de n de manera no lineal y los $\bar{p}_i$ denotan las proporciones de sentenciados estimados en cada EP i sobre la base del censo del 2016. Dado que no existe una rutina estándar de programación no-lineal entera bajo restricciones, optaremos por resolver (4.10) para cada posible valor entero de $n \in \{2, 3, \ldots, [\frac{C_0}{c_1}]\}$ y elegir luego el tamaño de muestra n de primera etapa, como el valor que minimice las varianzas de estas soluciones. Para esto usaremos el paquete `nloptr`(Ypma et al., 2018) de R, el cual es una interfase para resolver problemas de optimización con restricciones. Las restricciones de costos y opciones de optimización se programan en

```
gm <-function(m,n,M,pEPs,cc){ N = length(M)
c2 = cc[2:(N+1)]
pik = inclusionprobabilities(M,n)
sum(c2*pik*m) - (cc[N+2]-cc[1]*n)}
opts = list("algorithm"="NLOPT_LN_COBYLA","xtol_rel"=1.0e-8,maxeval = 2000)
```

y la función a minimizar se encuentra en

```
moptimn <-function(n,M,pEPs,cc){ N = length(M)
c2 = cc[2:(N+1)]
pik = inclusionprobabilities(M,n)
m0 = (cc[N+2]-cc[1]*n)/sum(c2*pik*pEPs*(1-pEPs))
ini = m0*pEPs*(1-pEPs)
ind = which(ini > M)
ini[ind] = M[ind]
ff = nloptr(x0 = ini,eval_f=Vem, lb=rep(0.0001,N),ub=as.numeric(M),
eval_g_ineq =gm, opts=opts,n=n,M=M,pEPs=pEPs,cc=cc)}
```

mientras que la gráfica de esta función para diferentes valores de n, obtenida con el código abajo mostrado, se aprecia en la figura 4.2.

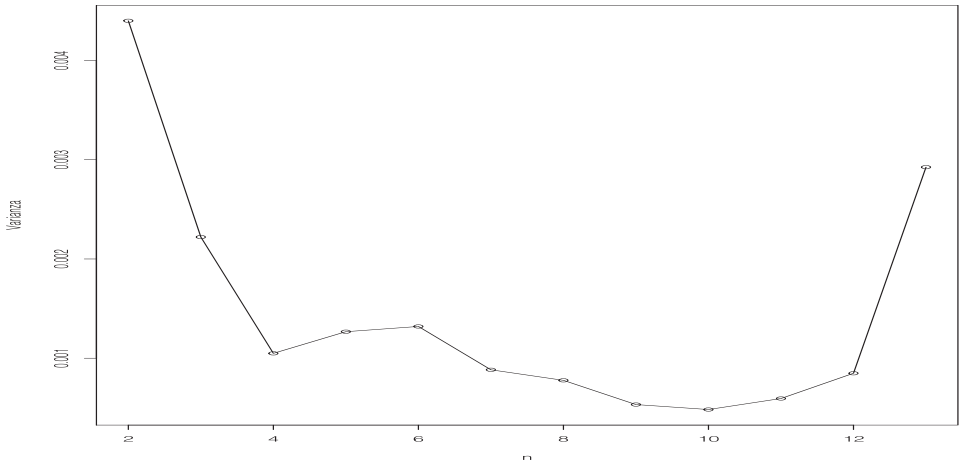


Figura 4.2: Varianza (4.6) de la proporción de sentenciados estimados para cada valor de n

```
v = 0
top = floor(cc[N+2]/cc[1])
for (h in 2:top){ aux = moptimn(h,M,pEPs,cc)
        v[h] = aux$objective}
v = v[-1]
plot(2:top,v,xlab="n",ylab="Varianza")
lines(2:top,v)
```

El tamaño de muestra recomendado será entonces de 10 EP y la cantidad de internos que se seleccionará en cada EP, en caso de que este sea elegido, se obtendrá de

```
opts = list("algorithm"="NLOPT_LN_COBYLA","xtol_rel"=1.0e-8,maxeval = 10000)
mm = moptimn(10,M,pEPs,cc)
round(mm$solution)

##  [1] 56 68 11 54 60 50 41 73 42 38 30 64 84 77 77 13 98 82 47 53 57 59 56
## [24] 62 52 52 42 31 75 61 70 49 62 57 56 63 38 60 57 54 49 57 60 13 41 44
## [47] 48 37 49 56 50 41 45 42 46 34 52 49 55 53 60 23 49 22 55 57 60 14 58
## [70] 62 60 18 58 43 60 49 44 17 59 10 55 10 58 33 46 29 57 51
```

4.15. Ejercicios

1. Un estudiante de un internado desea estimar el promedio final medio que alcanzaron él y sus compañeros en un curso de la institución. En lugar de obtener un listado de todos sus compañeros y realizar un MASs, se da cuenta de que los alumnos de su institución están distribuidos en 100 cuartos de 4 alumnos cada uno. Por ello decide seleccionar al azar 5 de estos cuartos y preguntarles a todos los estudiantes en esos cuartos el puntaje promedio que obtuvieron en el curso. Los resultados se muestran en la siguiente tabla:

Alumno	Cuarto				
N^o.	1	2	3	4	5
1	15.4	11.8	10	15	13.4
2	13	15.2	12.8	14.4	9.6
3	17.2	16.4	12.6	17.2	16.4
4	15.2	13.4	9.4	18.2	16

a) Obtenga la estimación buscada y su error estándar de estimación.
b) Obtenga un intervalo de confianza al 99 % para la estimación anterior.

2. En Richardson (2012) se presenta el mapa de la figura 5.3 que corresponde a un sitio arqueológico. Este contiene 100 cuadrículas de posible excavación, donde X denota a una cuadrícula que contiene artefactos o "hallazgos". Si usted tiene un presupuesto para seleccionar tan solo 20 cuadrículas, seleccione al azar una muestra de 20 cuadrículas siguiendo los diseños MASc, MASs, MAE con asignación proporcional (dividiendo el área en dos estratos I y II conformados, respectivamente, por las columnas 1-5 y 6-10), muestreo sistemático y muestreo por conglomerados de tamaño 2 (donde cada fila es un conglomerado). Para cada uno de los diseños estime el número total de cuadrículas que contendrán hallazgos en esta área. Indique en cada caso el margen de error cometido con un nivel de confianza del 95 %. Comente sus resultados.

1	2	3	4	5	6	7	8 X	9	10
11	12	13 X	14	15	16	17	18	19 X	20 X
21	22	23	24	25 X	26	27	28	29	30
31	32	33	34	35 X	36	37	38	39	40
41	42	43	44 X	45	46	47 X	48 X	49	50
51 X	52	53 X	54 X	55 X	56	57	58 X	59	60
61	62	63	64	65	66 X	67	68	69	70
71	72	73	74 X	75 X	76	77	78	79 X	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98 X	99	100 X

Figura 4.3: Mapa de un sitio arqueológico

3. A fin de estimar la proporción de poseedores de al menos un auto entre los 3000 empleados de una compañía que se divide en 20 departamentos de 150 funcionarios cada uno, se plantea un diseño que seleccionará al azar 10 departamentos y dentro de cada departamento 10 empleados. Si el número encontrado de empleados que poseen a menos un auto en esta muestra fue de $4, 5, 9, 0, 9, 9, 8, 6, 5, 4$, estime la proporción pedida y construya un intervalo de confianza al 95 % para este parámetro.

4. Una empresa de investigación de mercados ideó un plan de muestreo para estimar las ventas semanales de un producto A en una área geográfica. La empresa decidió muestrear ciudades dentro del área y luego supermercados dentro de cada una de las ciudades. La medición de interés es el número de cajas vendidas del producto A en una semana específica. Cinco ciudades son muestreadas de entre las 20 del área. Usando los datos presentados en la tabla adjunta

Ciudad	Número de supermercados	Número de supermercados muestreados	Media muestral	Desviación estándar muestral
1	45	9	102	20
2	36	7	90	16
3	20	4	76	22
4	18	4	94	26
5	28	6	120	12

a) Estime las ventas medias de todos los supermercados en el área para la semana específica y su intervalo de confianza máximo al 95 %. ¿Es insesgado el estimador utilizado?
b) ¿Se tiene suficiente información para estimar el número total de cajas del producto A vendidas en todos los supermercados del área durante la semana? Si es así, explique cómo estimaría este total y establezca un límite para el error de estimación. Use un nivel de confianza del 95 %.

5. Considere la base de datos poblacional Province 91 vista en el ejercicio 15 del capítulo 2, en donde la variable de conglomeración Cluster agrupa a un conjunto de municipalidades geográficamente contiguas de la provincia en estudio. Supongamos ahora que deseamos realizar una encuesta por muestreo utilizando ya sea un diseño por conglomerados de una etapa o de dos etapas. En la primera se seleccionarán mediante un MASs tres conglomerados y en la segunda se realizará un MASs de 4 conglomerados y dentro de estos un MASs de dos municipalidades. Si es de interés estimar el número de personas desempleadas en la provincia,
a) Halle la estimación pedida bajo los dos esquemas de muestreo.
b) Asumiendo que cuenta con toda la información, obtenga los efectos de diseños de ambos esquemas e indique cuál sería más eficiente.
c) Asumiendo que no cuenta con toda la información, estime los efectos de diseño anteriores.

6. En este ejercicio, tomado de Mendenhall et al. (2007), una socióloga desea estimar el número total de jubilados que viven en una ciudad. La socióloga decide muestrear manzanas y después casas dentro de las manzanas. Se seleccionaron aleatoriamente 4 manzanas de entre 300 de la ciudad. Responda a las siguientes preguntas a partir de los datos presentados en la tabla que aparece a continuación

Manzana	Número de casas	Número de casas muestreadas	Número de residentes jubilados por casa
1	18	3	1, 0, 2
2	14	3	0, 3, 0
3	9	3	1, 1, 2
4	12	3	0, 1, 1

a) Estime el número total de residentes jubilados en la ciudad y su error estándar de estimación.
b) Estime el número promedio de residentes jubilados por casa y su error estándar de estimación.
c) ¿Puede estimar el número promedio de residentes por manzana? Si su respuesta es afirmativa, obtenga esta estimación y su error estándar de estimación.

7. Muestre que la correlación intraclase definida en la sección 4.6 para diseños bietápicos en las que las USM son todas de un mismo tamaño M puede escribirse como

$$\rho = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\mu)(y_{ik}-\mu)}{(NM-1)(M-1)\sigma^2},$$

donde μ y σ^2 son, respectivamente, la media y la varianza poblacionales de la variable de investigación y, N es el número de UPM e y_{ij} es el valor que toma esta variable en la j-ésima USM de la UPM i.

8. Complete la demostración del teorema 4.1. Más concretamente, muestre que el estimador de Sen-Yates-Grundy para la varianza del estimador del total de Horvitz-Thompson es insesgado.

9. Consideremos un muestreo trietápico que busca estimar la media de una variable `y` en la que las unidades muestrales son todas de igual tamaño. Suponga que se tomarán secuencialmente un MASs de n UPM, un MASs de m USM dentro de cada UPM y un MASs de q UTM (unidades terciarias de muestreo) dentro de cada USM.
a) Encuentre una fórmula para la varianza del estimador.
b) Halle los tamaños de muestra óptimos en el sentido de que con ellos se minimice la varianza anterior o el costo total de muestreo $C = c_0 + c_1 n + c_2 nm + c_3 nmq$, sujeto a que se fije uno de ellos. Aquí, c_0 es un costo fijo, c_1 el costo por UMP seleccionado, c_2 el costo por USM seleccionado y c_3 el costo por UTM seleccionado.

10. En una población de 4 personas se van a seleccionar sin reemplazamiento 2 con probabilidades no constantes. Se sabe que la probabilidad de que se seleccionen a las dos primeras personas es 0.2, que se seleccionen a la primera y tercera es la misma que se seleccionen a la primera y cuarta, siendo esta de 0.1, que se seleccionen a la segunda y cuarta es la misma que se seleccionen a la tercera y cuarta, siendo esta de 0.15 y, finalmente, que se seleccionen a la segunda y tercera personas es de 0.3.
a) Halle las probabilidades de inclusión de cada persona en la muestra.
b) Si la población estadística del número de hermanos y de estas cuatro personas es, respectivamente, $\mathcal{P}_y = \{2, 1, 5, 4\}$, tome bajo este diseño una muestra de tamaño 2 y estime el número total de hermanos de esta población. Obtenga también una estimación del error estándar de estimación correspondiente.

11. Demuestre, utilizando el estimador de Horvitz-Thompson, que la varianza del estimador de la media poblacional para un muestro por conglomerados bietápico está dada por la expresión (4.2).

12. Muestre que el estimador $\hat{\tau}_\psi$ definido en la sección 4.9 es un estimador insesgado del total poblacional. Pruebe también que la varianza de este estimador viene dada por

$$V(\hat{\tau}_\psi) = \frac{1}{n}\sum_{i=1}^{N} \psi_i \left(\frac{\tau_i}{\psi_i} - \tau\right)^2 + \frac{1}{n}\sum_{i=1}^{N} \frac{V(\hat{\tau}_{ij})}{\psi_i}$$

y que (4.7) es un estimador insesgado de esta varianza.

13. Para conocer el rendimiento escolar en los colegios de una zona (6 en total) se ha planeado seleccionar aleatoriamente y sin reemplazo 2 de estos colegios con probabilidades proporcionales al número de alumnos de los colegios y luego seleccionar al azar 30 alumnos de cada colegio con el fin de aplicarles una prueba de conocimientos. La distribución del número de estudiantes por colegio de la zona es

Colegio	A	B	C	D	E	F
Número de estudiantes	150	200	50	30	400	100

Si realizada la selección anterior salieron elegidos los colegios A y E con los siguientes resultados:

Colegio seleccionado	Media	Varianza
1	14.5	25.64
2	10.9	16.36

a) Estime, de manera insesgada, el rendimiento medio de esta zona junto con su error estándar de estimación.
b) Si alguien le objeta que debió considerar, para que el muestreo sea representativo, iguales probabilidades de selección, ¿qué le respondería?

14. El año pasado una plaga de roya afectó seriamente la producción de café en cierta zona de un país que agrupa a 15 unidades agropecuarias (UA), las cuales se ubican en dos zonas ecológicas (1 = Baja y 2 = Alta). Con el objetivo de estimar las pérdidas medias en miles de dólares (μ) para los productores de café de esta zona a causa de la plaga, el ministerio del sector está interesado en realizar un estudio en la zona. Los datos siguientes ilustran la variable de pérdida en miles de soles (y), la variedad cultivada de café (A o B), el número de hectáreas (Ha) y las variables anteriormente descritas para cada unidad agropecuaria de la zona. Naturalmente, y se desconoce, pero se la presenta aquí solo para evitar que usted tenga que recabar esta información en el campo.

UA	Zona	Cooperativa	Ha	Variedad	y
1	1	1	41.5	A	7.3
2	1	1	23.8	A	6.2
3	1	1	33.3	B	7.2
4	1	1	22.1	A	4.8
5	1	2	44.8	A	7.6
6	1	2	37.3	A	7.4
7	1	2	29.5	A	5
8	1	3	21.5	B	5.8
9	1	3	18.4	B	2.2
10	1	3	13.7	A	6.1
11	2	4	12.5	B	4.5
12	2	4	15.2	B	4.8
13	2	5	6.5	B	2.8
14	2	5	5.8	B	3.3
15	2	5	10.4	A	5.1

a) Suponga que se seleccionarán mediante un MASs las UA 4, 9, 11 y 15 a fin de estimar μ. Halle el error estándar de estimación estimado respectivo.
b) Use los números aleatorios 0.231, 0.627, 0.122 y 0.883 para seleccionar, mediante un MASs, 4 UA. Estime con ello μ.
c) Asumiendo que conoce la tabla de arriba, halle bajo un MASs la desviación estándar de cualquier media muestral de tamaño 4 y estímela usando la muestra en b).
d) Tome un MAE con asignación proporcional y tamaño $n = 6$, con zona como variable de estratificación y estime μ y la proporción de UA en la región que cultivaron la variedad A. Use para ello el siguiente orden en la selección de UA de la zona 1 (hasta la cantidad necesaria): 3, 10, 5, 2, 9, 1, 4, 8, 6, 7 y el siguiente orden para la zona 2: 14, 15, 11, 13 y 12.
e) Suponga que al seguir el diseño en d) encontró que el muestreo por UA en la zona 1 cuesta aproximadamente 54 soles; mientras que en la zona 2 cuesta 40 soles, ¿cómo sugeriría para un estudio futuro distribuir la muestra de 6 UA de tal manera que minimice estos costos de muestreo? Use las estimaciones de d).

f) Suponga ahora que se aplica un muestreo por conglomerados de una etapa, siendo la variable de conglomeración la cooperativa. Si en este diseño se seleccionan al azar las cooperativas 1 y 4, estime μ bajo dos escenarios: uno en el que conozca el número de UA por cada cooperativa y otro en el que desconozca este número y lo averigue en el trabajo de campo.
g) De algún indicador en f) que le permita comparar este diseño con el MAE aplicado en d) y haga la comparación respectiva, indicando cuál de los diseños es más eficiente.
h) Si se tomará una muestra de 2 conglomerados (cooperativas) bajo un esquema sistemático ppt con tamaños proporcionales al número de hectáreas que administra cada cooperativa, ¿con qué probabilidad las cooperativas 1 y 4 serían seleccionadas?
i) Estime μ, bajo el esquema en h). Use el número aleatorio 0.305.

15. Realice, para el ejemplo de las ventas del supermercado, un pequeño estudio de simulación a fin de comprobar que el método de Sampford "funciona". Para ello, escriba un programa en R que seleccione 1000 muestras de tamaño 3 bajo este esquema y, con estas simulaciones, estime las probabilidades de inclusión ppt de primer orden. Compare luego estas con las verdaderas probabilidades ppt del ejemplo.

16. En el siguiente ejercicio, tomado de Mendenhall et al. (2007), un parque de diversiones cobra entrada por auto en lugar de por persona y desea estimar el número promedio de personas por auto en un día festivo. El funcionario del parque sabe por experiencia que entrarán a este alrededor de 400 autos y decide muestrear 80 de ellos. Para obtener una estimación de la varianza, decide utilizar un muestreo sistemático repetido con 10 muestras de 8 autos cada una. Usando los datos que a continuación se presentan, estime el número medio de personas por auto y establezca un límite para el error de estimación.

Inicio aleatorio	Segundo elemento	Tercer elemento	Cuarto elemento	Quinto elemento	Sexto elemento	Séptimo elemento	Octavo elemento
2(3)	52(4)	102(5)	152(3)	202(6)	252(1)	302(4)	352(4)
5(5)	55(3)	105(4)	155(2)	205(4)	255(2)	305(3)	355(4)
7(2)	57(4)	107(6)	157(2)	207(3)	257(2)	307(1)	357(3)
13(6)	63(4)	113(6)	163(7)	213(2)	263(3)	313(2)	363(7)
26(4)	76(5)	126(7)	176(4)	226(2)	276(6)	326(2)	376(6)
31(7)	81(6)	131(4)	181(4)	231(3)	281(6)	331(7)	381(5)
35(3)	85(3)	135(2)	185(3)	235(6)	285(5)	335(6)	385(8)
40(2)	90(6)	140(2)	190(5)	240(5)	290(4)	340(4)	390(5)
45(2)	95(6)	145(3)	195(6)	245(4)	295(4)	345(5)	395(4)
46(6)	96(5)	146(4)	196(6)	246(3)	296(3)	346(5)	396(3)

Las respuestas del número de personas por auto se encuentran entre paréntesis.

17. El organismo de medición de la calidad educativa de un país ideó un plan de muestreo para estimar el rendimiento medio de los alumnos del tercer año de educación secundaria de una región. El organismo decidió muestrear primero distritos educativos y luego colegios dentro de cada distrito. Cinco distritos son muestreados de entre los 15 de la región. Usando los datos que se muestran en el cuadro 4.15, donde se marca con X los distritos seleccionados,
a) Estime el rendimiento medio de los colegios en la región. ¿Es insesgado el estimador usado?
b) Obtenga un límite para el máximo margen de error en la estimación anterior al 95 %.
c) Suponga que en un estudio futuro se seleccionarán al azar tres distritos con probabilidades proporcionales al número de colegios en el distrito a fin de medir el impacto de un nueva política educativa para la región. Haga la selección y diga a qué distritos habría que hacerles el seguimiento en este estudio. Calcule también la probabilidad de que el distrito con un mayor número de colegios de la región participe de este estudio.

	Distrito escolar	Número de colegios	Número de colegios elegidos	Número de colegios unidocentes	Media	DE
X	1	25	9	3	15.25	3.06
	2	16		4		
	3	32		11		
X	4	26	7	3	13.56	2.18
	5	24		2		
	6	20		5		
	7	26		4		
	8	18		2		
X	9	30	4	6	12.17	2.45
	10	36		9		
X	11	28	4	5	10.65	2.60
	12	22		9		
	13	45		10		
X	14	39	6	8	15.38	2.93
	15	26		7		

Cuadro 4.3: Datos de la muestra para el ejercicio 17

18. Suponga que en la pregunta anterior se hubiese tenido interés en estimar la proporción de colegios unidocentes de la región y que con este fin se plantearan dos propuestas: seleccionar 4 distritos con probabilidades proporcionales al número de colegios en el distrito o seleccionar 4 distritos mediante un muestreo por conglomerados de una etapa.
a) Utilizando un esquema sistemático ppt para la primera propuesta, ¿sería posible reportar la proporción estimada buscada y su error estándar de estimación estimado?

b) Realice la selección de los 4 distritos y estime la proporción de colegios unidocentes en la región bajo las dos propuestas. Para la selección ppt use un esquema de Sampford.
c) A un nivel de confianza del 95 %, ¿qué error reportaría en sus estimaciones anteriores?. ¿Cuál propuesta consideraría que es la mejor?

19. Una cadena tiene 16 tiendas en el país. Si bien la cadena sugiere un precio de venta de 750 soles para un nuevo modelo de celular YTRON que llegó el mes pasado, este precio es variable y se deja a criterio del vendedor siempre que no sea inferior a los 680 soles, que es el precio de costo. Para estimar la proporción de celulares YTRON vendidos con rebaja sobre el precio ofrecido y estimar el monto total recabado hasta el momento por la venta de estos celulares, se piensa tomar una muestra de 4 de estas tiendas.
a) Si la muestra se tomó mediante un MASs y se obtuvieron los siguientes resultados:

Número de celulares YTRON en stock	Número de celulares YTRON vendidos	Monto total de ventas del celular YTRON	Número de celulares YTRON vendidos con rebaja
30	5	3730	1
45	10	7200	8
18	8	5670	6
20	9	7000	3

Reporte las estimaciones pedidas y sus errores estándar de estimación estimados.

b) Si la distribución del número de celulares YTRON destinados a cada tienda a inicios del mes (stock) fue la siguiente y la muestra se toma con probabilidades proporcionales al stock:

Tienda	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Stock	55	45	10	12	10	120	18	20	35	45	10	36	30	27	15	50

obtenga estas probabilidades y tome la muestra respectiva mediante un esquema sistemático ppt. Reporte la semilla aleatoria utilizada.
c) Suponga ahora que usted considera utilizar el estimador de Hansen-Hurwitz con reemplazamiento. Tome la muestra de 4 tiendas e indique cómo obtendría la estimación del total actual de ventas de los celulares YTRON en la cadena.

20. Suponga que para la ECE 2018 de la DRE Amazonas se le pide hacer un estudio de simulación que consiste en seleccionar 500 muestras de 50 colegios, cada una con probabilidades proporcionales al número de alumnos por colegio. Basándose en los esquemas de Poisson,sistemático simple y aleatorio, Tillé, Midzuno, pivotal y de conglomerados de una etapa, estime el rendimiento medio en Matemáticas. Indique a partir de un diagrama de cajas, cuáles de los esquemas anteriores logran una mayor precisión.

21. Suponga que en el ejemplo 4.6 se plantea un plan con las siguientes características:

- Las muestras se tomarán secuencialmente.
- La primera selección se tomará con probabilidad proporcional al área del supermercado, la segunda y tercera selección se harán al azar y con reemplazamiento, sin tomar en cuenta la primera selección.

Defina formalmente el diseño de muestreo asociado a este algoritmo.

22. Un ingeniero ambiental desea estimar el número total de árboles en un determinado condado que han sido afectados por una enfermedad y cuál es el nivel de esta infección. Hay 15 zonas forestales bien definidas en el condado, las cuales están divididas en parcelas de aproximadamente el mismo tamaño. Cuatro equipos están disponibles para el estudio, el cual deberá completarse en un día. Con este propósito se diseñó un muestreo aleatorio por conglomerados bietápico. En el que se seleccionaron al azar y sin reemplazamiento 4 zonas y 6 parcelas. Los datos recopilados del número de árboles afectados y entre paréntesis de la cantidad de ellos que tienen una infección avanzada se muestran en la siguiente tabla:

Zona	Número de parcelas	Número de árboles infectados (con infección avanzada)
1	12	15(5), 14(2), 21(8), 18(3), 9(1), 10(0)
2	16	4(0), 7(2), 10(1), 9(1), 8(3), 5(0)
3	14	10(3), 11(2), 14(2), 10(1), 9(0), 15(4)
4	21	6(2), 3(1), 4(1), 1(0), 2(0), 5(1)

a) Calcule la probabilidad de que una parcela particular de la zona 2 sea seleccionada.
b) Estime la proporción de árboles en la zona 1 que tienen una infección avanzada.
c) Estime la proporción de árboles en el condado que tienen una infección avanzada.
d) Suponga que la distribución del área en metros cuadrados de las parcelas y el número de árboles que estas contienen es, para la zona 2, la siguiente:

Parcela	1	2	3	4	5	6	7	8
Área	400	580	674	920	180	300	380	555
Num. de árboles	16	21	18	24	24	23	25	51
Parcela	9	10	11	12	13	14	15	16
Área	990	602	508	210	350	678	440	735
Num. de árboles	42	19	11	10	36	21	37	12

Si usted toma una muestra de 8 parcelas con probabilidades proporcionales al área de estas, ¿cuál sería su estimación y cuál su error estándar de estimación estimado para el número de árboles que contendría esta zona? Use para su muestreo un esquema de Sampford.

23. Suponga que en el ejemplo 4.6, no es ya de interés tomar un muestreo ppt, sino considerar iguales probabilidades de selección con excepción del supermercado D, el cual debe tener el doble de probabilidad de ser seleccionado que los otros supermercados.
a) ¿Cuáles serían las probabilidades de inclusión de primer orden bajo este esquema?
b) Halle el estimador de Horvitz-Thompson para el total de ventas en la cadena si salieron seleccionados, bajo este esquema, los supermercados A, D y E.
c) Si se seleccionan ahora al azar y sin reemplazamiento uno por uno cada uno de los 3 supermercados, utilizando en cada selección probabilidades proporcionales al tamaño, ¿con qué probabilidad será seleccionado el supermercado D?
d) Estime el total de ventas de la cadena, si se decide que en caso salga seleccionado el supermercado D, se medirá el total de ventas en 2 de sus 5 divisiones seleccionadas al azar. Suponga que el total de ventas en estas divisiones están en el orden de los 40, 45, 68, 29 y 63 miles de dólares.

Capítulo 5

Una introducción al muestreo complejo

La gran mayoría de encuestas por muestreo sobre poblaciones grandes involucran varias de las ideas analizadas: una encuesta puede estar segmentada en dominios, estratificada con varias etapas de formación de conglomerados, las probabilidades de selección pueden no ser iguales y es factible utilizar un muestreo sistemático en cualquiera de las etapas. Generalmente, la estratificación forma la clasificación más gruesa, los estratos pudieran ser áreas del país o tipos de habitat. Se extraen de los estratos muestras de conglomerados (a veces con varias etapas) y puede haber una posestratificación o interés a posteriori sobre algunos dominios. Todo esto hace, como se comprenderá, que las fórmulas para los errores de estimación en este tipo de diseños sean prácticamente inmanejables. En este capítulo, presentaremos una introducción a la obtención de estimadores y de sus varianzas en estos tipos de diseños. Comenzaremos analizando los pesos de muestreo, el cálculo de estimadores mediante estos pesos y la estimación de las varianzas de estos estimadores. Finalmente, brindaremos una introducción al análisis estadístico bajo muestras complejas.

Para tener una idea de la magnitud de los problemas comentados, consideremos el censo penitenciario 2016 como base de una encuesta futura por muestreo. Dadas las características y el tamaño de la población, pueden plantearse aquí varios diseños, uno de los cuales expusimos al término del capítulo anterior. Aun cuando los diseños clásicos estudiados son teóricamente factibles, en la práctica estos son inviables dadas las restricciones de costos y la complejidad de la logística subyacente. Una propuesta más realista para los penales podría ser, por ejemplo, optar por un diseño estratificado y por conglomerados bietápico. De manera natural, los estratos pudieran estar definidos, como en el capítulo 3, por el género y nivel de hacinamiento de las cárceles, las unidades primarias de muestreo (UPM) en cada estrato podrían tomarse como los establecimientos penitenciarios (EP) y, finalmente, las unidades secundarias de muestreo (USM) podrían ser los internos al interior de cada EP. Se puede también pensar en tres etapas, si previamente a la selección de los internos se seleccionan

al azar algunos pabellones. Otro punto que considerar es si se toman o no dominios en el estudio. Estos podrían estar constituidos por las oficinas regionales que tienen a su cargo la administración de un grupo de EP. Finalmente, no es necesario que en las distintas etapas se tome un MASs. En varias encuestas similares sobre cárceles de la región se han considerado muestreos sistemáticos en algunas de las etapas de selección (dada su simplicidad y logística). Más aún, dada la inequidad de los tamaños de los EP, podría resultar conveniente que en una o más de las etapas se realicen muestreos ppt.

5.1. Pesos de muestreo

El peso base de muestreo para una unidad de observación se define como el inverso de su probabilidad de selección. En un muestreo complejo, estos pesos requieren con frecuencia algunos ajustes adicionales por motivos, tales como la elegibilidad desconocida, la no respuesta y el uso de data auxiliar tendiente a reducir la varianza o corregir deficiencias en el marco muestral. En su forma definitiva, los pesos contienen prácticamente toda la información necesaria para construir un estimador puntual. Nosotros agregaremos un supraíndice 0 a estos pesos para enfatizar que son los pesos base.

Consideremos, por simplicidad, que nuestro interés sea estimar un total poblacional τ de una variable estadística y en una población de tamaño N. Entonces, el estimador puntual de τ tendrá la forma

$$\hat{\tau} = \sum_k \omega_k^0 y_k \delta_k \ ,$$

donde la suma va sobre todas las unidades de la población y las múltiples etapas de selección; δ_k es una v.a. indicadora de si la unidad k es seleccionada o no en la muestra, y los ω_k^0 son los pesos bases asociados a la selección de la unidad correspondiente a la medición y_k. Este estimador puede escribirse alternativamente como

$$\hat{\tau} = \sum_{i \in \mathcal{S}} \omega_i^0 Y_i,$$

donde la suma va sobre las unidades seleccionadas en la muestra bajo el diseño (que denotaremos por $\mathcal{S}$ y que es un subconjunto de la población $\mathcal{P}$) e Y_i denota a la v.a. correspondiente al valor que y toma en la i-ésima selección. Veamos algunos ejemplos y por ser más breve apelemos por ahora a la primera notación.

- En el MAS se tiene que

$$\hat{\tau} = \sum_{i=1}^{N} \omega_i^0 y_i \delta_i,$$

donde $\omega_i^0 = \frac{N}{n}$ es el inverso de la probabilidad de selección. Puesto que la suma de los pesos de las unidades seleccionadas es N, el estimador natural de la media poblacional

es μ; es decir, $\bar{Y}$ puede escribirse como

$$\bar{Y} = \frac{\sum_{i=1}^{N} \omega_i^0 y_i \delta_i}{\sum_{i=1}^{N} \omega_i^0 \delta_i}.$$

- En un MAE se tiene que

$$\hat{\tau} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \omega_{ih}^0 y_{ih} \delta_{ih},$$

donde $\omega_{ih}^0 = \frac{N_h}{n_h} = \frac{1}{P(\delta_{ih}=1)}$. Recordemos que por el hecho de que la suma de los pesos de las unidades seleccionadas sea N, cada unidad en la muestra "representa" cierta cantidad de unidades de la población de modo que toda la muestra "representa" la población. La estimación de la media para el muestreo estratificado es

$$\bar{Y} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{N_h} \omega_{ih}^0 y_{ih} \delta_{ih}}{\sum_{h=1}^{H} \sum_{i=1}^{N_h} \omega_{ih}^0 \delta_{ih}}.$$

- En un muestreo por conglomerados bietápico se tiene que

$$\hat{\tau} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \omega_{ij}^0 y_{ij} \delta_{ij},$$

donde $\omega_{ij}^0 = \frac{NM_i}{nn_i}$, y la estimación de la media poblacional es

$$\bar{Y} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \omega_{ij}^0 y_{ij} \delta_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \omega_{ij}^0 \delta_{ij}}.$$

Es interesante notar que todos estos estimadores del total son por construcción insesgados y que tales esquemas pueden utilizarse para obtener el estimador de un total en otros diseños complejos. Consideremos, por ejemplo, el caso de un muestreo por conglomerados de tres etapas o trietápico. Aquí, la probabilidad conjunta de que la unidad terciaria k, de la unidad secundaria j perteneciente a la unidad primaria i sea seleccionada, puede calcularse por

$$\pi_{ijk} = P(\delta_{ijk} = 1) = \text{P(Seleccionar la unidad } k \mid \text{se seleccionaron las unidades } i \text{ y } j)$$

$$\times \text{ P(Seleccionar la unidad j} \mid \text{se seleccionó la unidad i)} \times \text{P(Seleccionar la unidad } i).$$

Luego, el peso de muestreo para esta unidad de observación viene dada por

$$\omega_{ijk}^0 = \omega_{k|i,j}^0 \times \omega_{j|i}^0 \times \omega_i^0,$$

siendo, respectivamente, $\omega_{k|i,j}^0$, $\omega_{j|i}^0$ y ω_i^0 los inversos de las probabilidades arriba indicadas.

5.1.1. Ajuste de pesos por no respuesta

Hasta el momento hemos implícitamente asumido que contamos siempre con un marco muestral perfecto y que toda unidad seleccionada en la muestra ha de responder a la encuesta o al instrumento de recolección de información. En la práctica, como es de esperarse, esto raramente ocurre, lo cual origina errores de no muestreo tanto en la cobertura como en la no respuesta. Una manera de mitigar estos errores es efectuando algunos ajustes a los pesos base (otra alternativa sería la imputación).

Supongamos que deseamos calcular la media μ de una variable estadística y en una población de tamaño N a la cual subdividiremos en dos grupos de tamaños N_r y N_m de medias μ_r y μ_m para y. Estas subpoblaciones incluyen, respectivamente, a los que responden y a los que no en la encuesta sobre la variable y. Puesto que solo μ_r podría conocerse (bajo un censo), el sesgo que uno cometería al reportar esta media como la de la población vendrá dado por

$$\mu_r - \mu = \mu_r - (\frac{N_r}{N}\mu_r + \frac{N_m}{N}\mu_m) = \frac{N_m}{N}(\mu_r - \mu_m).$$

Así incurriremos en un mayor sesgo mientras la proporción de no respuesta sea más grande (o de respuesta menor) o las medias de y para los que responden y no difieran más. Tomada la muestra, este sesgo podría teóricamente estimarse por

$$\frac{n_m}{n}(\bar{Y}_r - \bar{Y}_m),$$

donde n_m es el número de unidades sin respuesta en la muestra e $\bar{Y}_r$ y $\bar{Y}_m$ son, respectivamente, las medias muestrales de y para los que responden y no. La idea de una reponderación o ajuste es tratar de que $\bar{Y}_r$ e $\bar{Y}_m$ sean los más parecidos posibles, tarea ciertamente complicada, pues en la práctica uno no conoce $\bar{Y}_m$ ni, a priori, la proporción de unidades que han de responder a la encuesta.

Antes de entrar propiamente en la ponderación, es bueno entender cómo se podría generar una no respuesta y como, según ello, se podría especificar un elemento vital en todo este análisis: la probabilidad ϕ_k de que una unidad k seleccionada responda. Para ello seguiremos la terminología dada por Little y Rubin (2002), quienes suponen un modelo para el vector de variables de interés. Supongamos que en una encuesta tenemos para cada unidad k un vector de variables de interés $\mathbf{y}_k$ disponible solo si k responde y un vector de variables auxiliares $\mathbf{x}_k$ siempre disponible al margen de si la unidad k responde o no. Diremos que una no repuesta será:

- MCAR (de Missing Completely at Random). Si la probabilidad de respuesta para la unidad k, ϕ_k, no depende de $\mathbf{y}_k$ ni de $\mathbf{x}_k$. Ello ocurriría, por ejemplo, cuando al modelar la probabilidad ϕ_k de respuesta para los distintos elementos, estas resulten ser aproximadamente iguales.

- MAR (Missing at Random). Si la probabilidad de respuesta para la unidad k, ϕ_k, depende de todas o algunas de las variables auxiliares $\mathbf{x}_k$.

- NINR (Nonignorable Nonresponse). Si la probabilidad de respuesta para la unidad k, ϕ_k, depende de todas o algunas de las variables de interés $\mathbf{y}_k$ y esta dependencia no puede ser removida con un modelamiento sobre las $\mathbf{x}_k$.

Supongamos ahora que deseamos estimar, bajo un diseño complejo, un total para una variable estadística y en una población de tamaño N. Con el fin de incorporar la posibilidad de no respuesta, definamos una variable aleatoria indicadora R_k que vale 1 si, y solamente si, esta unidad k responde condicionada, claro está, a que sea seleccionada. En caso contrario, R_k vale 0. Tomada la muestra, y considerando solo las unidades con respuesta, un estimador de τ tiene la forma

$$\hat{\tau} = \sum_k \omega_k y_k \delta_k R_k.$$

El valor esperado de este estimador puede calcularse por

$$E(\hat{\tau}) = E(E(\hat{\tau} \mid \boldsymbol{\delta})) = E(\sum_k \omega_k y_k \delta_k E(R_k \mid \boldsymbol{\delta}))$$

$$= \sum_k \omega_k y_k E(\delta_k)\phi_k = \sum_k \omega_k y_k \pi_k \phi_k.$$

Así, este estimador será insesgado si consideramos pesos iguales a

$$\omega_k = \frac{1}{\pi_k \phi_k} = \omega_k^0 \frac{1}{\phi_k},$$

donde ω_k^0 es el peso base de muestreo para la unidad k.

La obtención de los pesos últimos implica, entonces, estimar las probabilidades de respuesta para cada unidad seleccionada, ϕ_k. Si asumimos que las no respuestas son MCAR o MAR, estas probabilidades podrían estimarse identificando alguna o algunas variables auxiliares bajo cuyos niveles se pueda predecir si la unidad k ha de responder o no. Ello puede hacerse con cualquier técnica de clasificación, como, por ejemplo, la regresión logística binaria. Si bien esto nos conducirá a una estimación probablemente distinta para cada unidad, en la práctica se aconseja ajustar por grupos de unidades. Estos pueden formarse, por ejemplo, si usamos una regresión binaria, ordenándose las probabilidades estimadas ϕ_k y clasificándolas mediante cuantiles. Luego podríamos ajustar los pesos bases de cada grupo con, por ejemplo, la inversa del promedio de las probabilidades ϕ_k dentro de cada grupo. Un ejemplo de la aplicación de esta técnica, puede verse en el ejercicio 5.4.

5.1.2. Ajuste de pesos por elegibilidad desconocida

Por más depuración hecha al marco muestral, es posible que este aún contenga unidades cuya elegibilidad no pueda predeterminarse. Esto es, unidades que no son posibles de contactar en la encuesta y, por tanto, su respuesta será incierta. Al igual que en el ajuste por no respuesta, el ajuste por elegibilidad desconocida se hace con las mismas clases que en esta y simplemente consiste en multiplicar el peso base por el cociente entre la suma de los pesos base de la clase dividida entre la suma de los pesos base de las unidades en la clase cuya elegibilidad sea conocida (sea que ellas respondan o no a la encuesta).

5.2. Estimadores no lineales

Si bien el uso de los pesos resuelve el problema de encontrar estimadores puntuales de totales, medias o proporciones en un diseño complejo, ello no nos da información acerca de la forma de determinar sus errores estándar. Las varianzas de los estimadores dependen de las probabilidades de que cualquier pareja de unidades sea seleccionada para estar en la muestra y requieren más conocimiento del diseño que el dado simplemente por los pesos.

Otro problema que surge con esta metodología es que, en apariencia, ella está restringida a la estimación de totales, medias o proporciones y no cubre a otros parámetros que podrían ser de interés, tales como medianas, desviaciones estándar, cuantiles, correlaciones u otros. Veremos a continuación que tal idea no es del todo cierta y que sí es posible estimar estas (más no directamente sus errores estándar) sobre la base de los pesos de muestreo. Si N es el tamaño de la población, la idea es aproximar con los pesos la verdadera proporción y la verdadera proporción acumulada poblacional de los valores de la variable de interés y. Estas vienen dadas respectivamente por

$$P(y) = \frac{\text{número de unidades cuyo valor es y}}{N}$$

y

$$F(y) = \frac{\text{número de unidades cuyo valor } \leq \text{ y}}{N} = \sum_{x \leq y} P(x).$$

Para ello definiremos, basándonos solo en la muestra, la función de probabilidad empírica

$$\hat{P}(y) = \frac{\sum_k \omega_k \mathbf{1}_{y_k = y} \delta_k}{\sum_k \omega_k \delta_k}$$

y su función de distribución empírica

$$\hat{F}(y) = \sum_{x \leq y} \hat{P}(x).$$

En otras palabras, $\hat{P}(y)$ es igual a la suma de los pesos de todas las observaciones en la muestra que toman el valor y, dividida entre la suma de todos los pesos en la muestra; y,

por otro lado, $\hat{F}(y)$ es la suma de los pesos para todas las observaciones en la muestra con valores menores o iguales que y, divididas entre la suma de todos los pesos en la muestra.

Si ahora deseamos estimar ciertos parámetros poblacionales, deberemos, en primer lugar, expresar estos en términos de su real proporción poblacional; por ejemplo, la media y varianza se expresan respectivamente por

$$\mu = \sum_y yP(y) \qquad \text{y} \qquad \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2 = \frac{N}{N-1}(\sum_y y^2P(y) - \mu^2).$$

Hecho esto, la estimación procederá al sustituir $\hat{P}(y)$ o $\hat{F}(y)$ en cada aparición de $P(y)$ ó $F(y)$.

Ejemplo 5.1. *Considere, para la ECE 2019 de la DRE Amazonas, un diseño estratificado de conglomerados de una etapa, donde la variable de estratificación será la definida por el cruce de las variables de gestión y área, y los conglomerados serán los colegios. Nuestro interés recaerá, en primer lugar, en seleccionar una muestra de 20, 20, 4 y 4 colegios en, respectivamente, los estratos Urbano.Estatal, Rural.Estatal, Urbano.No estatal y Rural.No estatal. y, en segundo lugar, en analizar cómo hace el paquete survey para estimar el rendimiento medio en Ciencia y Tecnología, a partir de solo los pesos base de muestreo. Para lo primero usaremos el comando* `mstage`, *el cual exige ordenar la base de datos por la variable de estratificación. Los códigos son los siguientes:*

```
library(survey)
library(sampling)
load("ece19Am.RData")
Pop = ece19Am
Pop$Estrato=interaction(Pop$area,Pop$gestion2)
Pop = Pop[order(Pop$Estrato),]
set.seed(12345)
disl = list("stratified","cluster")
m=mstage(Pop,stage=disl,varnames=list("Estrato","ID_IE"),
size=list(size1=table(Pop$Estrato),size=c(20,20,4,4)),method=list("","srswor"))
mues = getdata(Pop,m)[[2]]
mues$w0 = 1/mues$Prob
aa = by(Pop$ID_IE,Pop$Estrato,unique)
aa = as.vector(unlist(lapply(aa,length)))
mues$fpc = rep(aa,table(mues$Estrato))
```

Note que a la base de datos muestral `mues` *le hemos agregado, los pesos base de muestreo* ω_0 *y el número de colegios por estrato* fpc. *Para estimar el rendimiento medio en Ciencia y*

Tecnología debemos definir el diseño correspondiente. Ello podría hacerse en R con cualquiera de los siguientes dos comandos:

```
(disc0=svydesign(ids=~ID_IE,strata=~Estrato,fpc= ~fpc,data=mues,nest=T))

## Stratified 1 - level Cluster Sampling design
## With (48) clusters.
## svydesign(ids = ~ID_IE, strata = ~Estrato, fpc = ~fpc, data = mues,
##     nest = T)

(disc1=svydesign(ids=~ID_IE,strata=~Estrato,data= mues,weights=~w0))

## Stratified 1 - level Cluster Sampling design (with replacement)
## With (48) clusters.
## svydesign(ids = ~ID_IE, strata = ~Estrato, data = mues, weights = ~w0)
```

La diferencia entre ambos es que el primero respeta estrictamente la forma en que se obtuvo la muestra; mientras que el segundo considera los mismos pesos de la primera, pero asume que cada selección de los colegios al interior de los estratos se hace mediante un MASc; es decir, con reemplazamiento. Dado que los pesos no cambian, ambos nos brindarán las mismas estimaciones (bajo el estimador de razón), pero no necesariamente los mismos errores estándar de estimación estimados. Cabe recordar que cada vez que se omite el factor de corrección para poblaciones finitas `fpc` *en* `svydesign`*, uno implícitamente está asumiendo un muestreo con reemplazamiento. Una pregunta de interés al respecto sería qué hacer si deseamos llevar a cabo un esquema sin reemplazamiento en el cual se conozcan los pesos de muestreo. El paquete* `survey` *permite esta posibilidad, pero para ello se deben de realizar ciertas aproximaciones o, en todo caso, debe proveerse al comando* `svydesign` *de las probabilidades de inclusión y de la matriz de probabilidades de inclusión de segundo orden.*

Viremos ahora, al otro objetivo de este ejemplo. La estimación del rendimento medio en Ciencia y Tecnología viene dada, por el paquete survey, por

```
coef(svymean(~M500_CT,disc1,na.rm=T))

## M500_CT
##     444
```

Esta estimación es obtenida, precisamente, a través de la función de distribución empírica y el código

```
h = by(mues$w0,mues$M500_CT,sum)
Phat = as.vector(h/sum(h))
(meanCT = sum(as.numeric(names(h))*Phat))

## [1] 444
```

□

Un tratamiento especial se da para el caso de la estimación del cuantil $p \in [0, 1]$,

$$q_p = \text{mín}\{y \ / \ F(y) \geq p\}.$$

Si bien podríamos sustituir directamente aquí $F(y)$ por $\hat{F}(y)$, resulta más conveniente utilizar en su lugar una interpolación lineal entre los valores muestrales que tengan una proporción acumulada cercana a p. Esto nos conlleva al siguiente estimador para el cuantil p:

$$\hat{q}_p = y_1 + \frac{p - \hat{F}(y_1)}{\hat{F}(y_2) - \hat{F}(y_1)}(y_2 - y_1),$$

donde y_1 es el mayor valor y en la muestra que satisfaga $\hat{F}(y) < p$, e y_2 es el menor valor y en la muestra que cumpla $\hat{F}(y) > p$.

Ejemplo 5.2. *Se desea implementar un programa para adultos mayores de una pequeña comunidad. El programa se brindará al cuarto superior de las personas de mayor edad, por lo cual es de interés estimar el cuantil 0.75 de esta población. Si suponemos que las edades de todos los habitantes de la comunidad, segmentados en distritos, es la que se muestra en el cuadro 5.1, tome un MASs 10 personas y luego realice un muestreo por conglomerados bietápico de dos distritos y 10 personas en estos, con un número de USM proporcionales al tamaño del distrito, a fin de estimar el cuantil requerido bajo ambos diseños. Realice estas estimaciones con su propia rutina y usando el comando* `svyquantile` *del paquete survey.*

Distrito	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B
Edad	20	66	46	61	53	69	50	12	64	46	48	11	38	8	62	51
Distrito	B	B	B	B	B	B	B	C	C	C	C	C	C	C	C	C
Edad	38	11	35	65	59	90	19	11	54	56	11	47	54	63	33	17
Distrito	C	C	C	C	C	C	C	D	D	D	D	D	D	D	D	D
Edad	72	67	34	47	10	23	52	17	12	20	31	12	48	3	34	37
Distrito	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D
Edad	1	6	28	11	36	2	10	45	1	10	51	11	18	57	23	17

Cuadro 5.1: Distritos de pertenencia y edades en años de todos los miembros de la comunidad del ejemplo 5.3

Solución: Luego de crear el data frame Eje3cap5 mediante

```
Distrito = c("A", "A","A","A","A","A","A","A","A","A","B","B","B","B","B",
"B","B","B","B","B","B","B","B","B","C","C","C","C","C","C","C","C","C","C",
"C","C","C","C","C","D","D","D","D","D","D","D","D","D","D","D","D","D","D",
"D","D","D","D","D","D","D","D","D","D","D")
Edad = c(20,66,46,61,53,69,50,12,64,46,48,11,38,8,62,51,38,11,35,65,59,90,
19,11,54,56,11,47,54,63,33,17,72,67,34,47,10,23,52,17,12,20,31,12,48,3,34,37,
1,6,28,11,36,2,10,45,1,10, 51,11,18,57,23,17)
Eje3cap5 = data.frame(Distrito=Distrito,Edad=Edad)
```

La estimación del cuantil buscado bajo un MAS se hará mediante

```
set.seed(12345)
N = dim(Eje3cap5)[1]
sampleMASs = Eje3cap5[sample(N,10),]
dise1 = svydesign(id=~1,fpc = rep(N,10),data = sampleMASs)
svyquantile(~Edad,dise1,0.75)

##      0.75
## Edad 45.5

quantile(Eje3cap5$Edad,0.75)

##  75%
## 52.2
```

Ella nos brinda una estimación bastante pobre del verdadero tercer cuartil que está entre 52 y 53 años. Por otro lado, para la estimación por el diseño bietápico, primero será necesario definir los pesos de muestreo. Como recordamos, estos serán el producto del peso para la primera etapa que es 2 por el peso para la segunda etapa que dependerá de los distritos elegidos. Por las condiciones dadas, los tamaños de muestra posibles para la segunda etapa los podremos calcular mediante

```
ms = combn(4,2,function(x){
        h = as.vector(table(Eje3cap5$Distrito))
        round(10*h[c(x[1],x[2])]/sum(h[c(x[1],x[2])]))})
ms

##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    4    4    3    5    4    4
## [2,]    6    6    7    5    6    6
```

Como se ve, estos son de 4 y 6 residentes casi siempre, salvo que se seleccionen los distritos A y D o los distritos B y C. Al realizar el muestreo, obtuvimos

```
set.seed(12345)
(s = sample(6,1))

## [1] 5

(m = ms[,s])

## [1] 4 6
```

los distritos B y D, donde cabe recordar que, en la medida de lo posible, estamos siempre utilizando la semilla aleatoria 12345 para efectos de reproductibilidad. Esto implica que el peso 2 de la primera etapa tendrá que multiplicarse por 3.5 para el distrito B y por 4.167 para el distrito D, quedando la muestra final y sus pesos dados por

```
set.seed(12345)
m1 = sample(which(Eje3cap5$Distrito=="B"),4)
m2 = sample(which(Eje3cap5$Distrito=="D"),6)
Muestra2 = cbind(Eje3cap5[c(m1,m2),],Peso = c(rep(7,4),rep(8.33,6)))
(Muestra2 = cbind(Muestra2,fpc1 = rep(4,10),fpc2 = c(rep(14,4),rep(25,6))))

##    Distrito Edad Peso fpc1 fpc2
## 21        B   59 7.00    4   14
## 22        B   90 7.00    4   14
## 20        B   65 7.00    4   14
## 23        B   19 7.00    4   14
## 51        D   28 8.33    4   25
## 43        D   31 8.33    4   25
## 47        D   34 8.33    4   25
## 64        D   17 8.33    4   25
## 55        D   10 8.33    4   25
## 59        D   51 8.33    4   25
```

La estimación pedida, que dejamos para que la trabaje manualmente como ejercicio, se obtendrá finalmente a través de

```
dise2 = svydesign(ids=~Distrito+Edad,fpc=~fpc1+fpc2,data=Muestra2)
svyquantile(~Edad,dise2,0.75)
```

```
##      0.75
## Edad 52.7
```

Ella, como se aprecia, nos da una mucho mejor estimación del tercer cuartil pedido. □

5.3. Efectos de diseño y consideraciones prácticas para obtener tamaños de muestra

Obtener tamaños de muestra en muestreo complejo es una labor complicada y pocas veces es posible obtener una formulación explícita. En tal situación, como lo sugirió Kish (1965), es mejor usar las estimaciones de los efectos de diseño.

Como recordamos, el efecto de diseño está definido como el cociente entre la varianza del estimador bajo un muestreo complejo y la varianza de este estimador bajo un muestreo aleatorio simple, que según nuestra convención es sin reemplazamiento. En el caso de la media, este efecto viene dado por

$$deff = \frac{V_{mc}(\bar{Y})}{(1-\frac{n}{N})\frac{\sigma^2}{n}},$$

donde V_{mc} denota la varianza del estimador bajo el muestreo complejo. En la práctica, este efecto se desconoce al depender de características poblacionales, pero es factible de estimarse mediante

$$\widehat{deff} = \frac{\hat{V}_{mc}(\bar{Y})}{(1-\frac{n}{\hat{N}})\frac{\hat{\sigma}^2}{n}}.$$

Claramente, obtener esta cantidad requiere de una estimación de la varianza del estimador bajo el muestreo complejo, punto que detallaremos en la siguiente sección. Será también necesario estimar la desviación estándar de la variable en estudio y. El problema con esta última es que nosotros no hacemos un MASs sino un muestreo complejo, por lo cual su estimación solo debe basarse en los pesos de muestreo de este último diseño. Por fortuna, contamos, como lo detallamos en la sección anterior, con una manera de estimar esta desviación estándar, $\hat{\sigma}$, basándonos solo en los datos del muestreo complejo. Si bien esta estimación es en teoría: $\hat{\sigma} = \sqrt{\frac{N}{N-1}(\sum_y y^2\hat{P}(y) - \hat{\mu}^2)}$, donde $\hat{\mu} = \sum_y y\hat{P}(y)$ y $\hat{P}$ denota la función de probabilidad empírica que se define solo con los pesos de muestreo del diseño complejo, el tamaño de la población N no siempre es conocido y de allí que se considere para este un estimador $\hat{N}$ en el deff. Este tamaño puede estimarse con la suma de los pesos de muestreo y podría no coincidir con N por los ajustes en los pesos. Además si las muestras son pequeñas, se conseguirán estimadores más fiables si reemplazamos el cociente $\frac{N}{N-1}$ por $\frac{n}{n-1}$. Esta es precisamente la metodología utilizada por el paquete survey de R para estimar los efectos de diseño. El siguiente ejemplo ilustra la estimación de este efecto.

Ejemplo 5.3. *Retomemos el ejemplo 5.1 y supongamos que nos piden estimar el efecto de diseño en la estimación del rendimiento medio en Ciencia y Tecnología. Los códigos del caso vienen dados por*

```
(mCT = svymean(~M500_CT,disc1,na.rm=T,deff=T))

##          mean    SE DEff
## M500_CT 443.9  15.5 22.1
```

o por

```
h = by(mues$w0,mues$M500_CT,sum)
Phat = as.vector(h/sum(h))
(meanCT = sum(as.numeric(names(h))*Phat))

## [1] 444

sum2 = sum(as.numeric(names(h))^2*Phat)
n = sum(is.na(mues$M500_CT)==0)
N = sum(h)
sigma2_e = (n/(n-1))*(sum2-meanCT^2)
(deff_e = (SE(mCT)^2)/((1 - n/N)*sigma2_e/n))

##         M500_CT
## M500_CT    22.1
```

□

Supongamos ahora, asumiendo que contamos con una estimación del efecto de diseño, que deseamos determinar el tamaño de muestra necesario n a utilizar en un muestro complejo, de tal manera que el error en la estimación de la media sea no mayor que e con un nivel de confianza de $100(1-\alpha)\,\%$; es decir:

$$e = z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}_{mc}(\bar{Y})}.$$

De la fórmula para estimar el efecto de diseño podríamos, entonces, despejar la estimación de la varianza $\hat{V}_{mc}(\bar{Y})$ y reemplazarla en esta última para obtener

$$e = z_{1-\frac{\alpha}{2}}\sqrt{\widehat{deff}\ (1-\frac{n}{\hat{N}})\frac{\hat{\sigma}^2}{n}}.$$

Así, despejando n de la ecuación

$$n = \frac{z^2_{1-\frac{\alpha}{2}}\widehat{deff}\ \hat{\sigma}^2\hat{N}}{\hat{N}e^2 + z^2_{1-\frac{\alpha}{2}}\widehat{deff}\ \hat{\sigma}^2}.$$

Note que si el tamaño de la población N o $\hat{N}$ es grande, se tendrá que aproximadamente

$$n = \widehat{deff}\ n_0,$$

donde n_0 es el tamaño de muestra para un MASs con poblaciones infinitas. En la práctica, el cálculo de estos tamaños de muestra debe aún corregirse ante la posibilidad de no respuestas. Las tasas de no respuestas tnr son fácilmente estimables de experiencias pasadas y se miden como la proporción de sujetos en estudio que no respondieron al estudio. Esta tasa obviamente incrementará el tamaño de muestra anterior y conllevará a un tamaño de muestra final igual a

$$n_f = \frac{n}{1 - \hat{tnr}},$$

siendo $\hat{tnr}$ la tasa de no respuesta estimada. Así, si se calculó $n = 500$ y se estima una tasa de no respuesta del 7 %, el tamaño de muestra final que debería considerarse es de $n_f = 538$ unidades.

El desarrollo hasta el momento descrito constituye la metodología más común para el cálculo de los tamaños de muestra en muestras complejas. El lector interesado puede indagar esto en diversos estudios. Un ejemplo se puede ver en

`http://observatorio.ministeriodesarrollosocial.gob.cl/layout/doc/casen/Informe%20Diseno%20Muestral_Revision_13sep12.pdf.`

Este es un informe que describe el diseño muestral de la Encuesta de Caracterización Socioeconómica Nacional (Casen) 2011 realizada en Chile.

Otro punto importante al planificar una muestra sobre una gran población es si se van a considerar dominios de estudio o no. Recordemos que los dominios conforman en general una partición de la población para las que se toman muestras independientes a fin de controlar la precisión de las estimaciones sobre cada dominio y sobre toda la población. Calculado el tamaño de muestra para la población, digamos n, una pregunta de interés sería saber cómo distribuir estos en los D dominios de estudio si es que estos existieran. Un criterio podría ser tomándolos de forma proporcional al tamaño N_d de cada dominio; sin embargo, ello podría resultar oneroso o producir estimaciones poco confiables en algunos de ellos, con márgenes de error superiores a los diseñados para el dominio.En su lugar, Bankier (1988) propuso minimizar alguna función criterio sobre el error relativo que se cometería bajo cierta asignación. Nosotros extenderemos esta idea considerando también la incorporación de costos unitarios c_d por selección en los dominios y el uso de efectos de diseño. Concretamente, si estamos interesados en estimar la media de una variable y para la población, buscaremos la asignación de la muestra total a los dominios que minimize la función

$$\sum_{d=1}^{D} (X_d^{\alpha} CV(\bar{Y}_d))^2 \tag{5.1}$$

o que minimize el costo total de muestreo, sujeto a la condición de que los tamaños de muestra por dominio n_d satisfagan la restricción $n = \sum_{d=1}^{D} n_d$. Aquí X_d denota la importancia del dominio d, que por lo usual es su tamaño, y $\alpha \in [0, 1]$ es un valor que queda a criterio del investigador y que modela la relevancia de la importancia que se le dé a cada dominio. Mientras α sea más pequeño, los dominios más pequeños, o de menor importancia, tenderán a tener una mejor representación. Un valor de compromiso es $\alpha = 0.5$.

El coeficiente de variación en (5.2) viene dado por

$$CV(\bar{Y}_d) = \frac{\sqrt{V(\bar{Y}_d)}}{\mu_d} \times 100,$$

donde μ_d es la media poblacional del dominio d y se asume que $\bar{Y}_d$ es un estimador insesgado de μ_d. El problema con este coeficiente es que si el muestreo es complejo la desviación estándar de $\bar{Y}_d$ es difícil de obtener, por lo cual podríamos usar los efectos de diseño $deff_d$ para los dominios a fin de reescribir (5.2) como

$$\begin{array}{ll} \text{mín} & \sum_{d=1}^{D} (\frac{X_d^{\alpha}\sqrt{deff_d}}{\mu_d})^2 (1 - \frac{n_d}{N_d}) \frac{\sigma_d^2}{n_d}, \\ \text{s.a.} & \sum_{d=1}^{D} n_d = n \end{array} \tag{5.2}$$

siendo σ_d^2 la varianza en el dominio d. Si se sigue literalmente la prueba de la proposición 3.2 (véase el ejercicio 13), no es difícil mostrar que la solución de (5.3) o de la minimización del costo total de muestreo, fijado un nivel para (5.2) o para el costo total, viene dado por

$$n_d = \frac{\frac{X_d^{\alpha}\sqrt{deff_d}}{\mu_d}\frac{\sigma_d}{\sqrt{c_d}}}{\sum_{j=1}^{D}\frac{X_j^{\alpha}\sqrt{deff_j}}{\mu_j}\frac{\sigma_j}{\sqrt{c_j}}} n.$$

En el caso particular de que los costos de muestreo y efectos de diseño por dominio sean los mismos, esta fórmula se simplifica a

$$n_d = \frac{\frac{X_d^{\alpha}\sigma_d}{\mu_d}}{\sum_{j=1}^{D}\frac{X_j^{\alpha}\sigma_j}{\mu_j}} n.$$

Obviamente, para calcular esta cantidad será necesario contar con estimaciones de los distintos parámetros y efectos de diseño por dominio, los cuales pueden provenir de algún estudio pasado o una muestra piloto.

5.4. Estimación de la varianza

Si bien los pesos de muestreo son de gran utilidad para incorporar el diseño en la obtención de la mayoría de estimaciones de interés, ellos no nos dicen nada acerca de sus varianzas y errores estándar de estimación. En esta sección abordaremos el problema de estimar estas varianzas; para ello en la literatura se han considerado dos enfoques: uno clásico de linealización y otro a través de métodos de remuestreo.

5.4.1. El método de linealización

Consideremos una población en la que nos interese estimar un parámetro θ expresable como una función suave (es decir, con derivadas continuas) de q totales de la población; vale decir,

$$\theta = h(\tau_1, \tau_2, \ldots, \tau_q).$$

Si $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_q$ son estimadores insesgados de, respectivamente, $\tau_1, \tau_2, \ldots, \tau_q$, entonces un estimador natural de θ viene dado por

$$\hat{\theta} = h(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_q).$$

Con el fin de encontrar la varianza de este estimador, podemos usar una expansión lineal basada en el teorema de Taylor y aproximar $\hat{\theta}$ alrededor de su verdadero valor θ mediante

$$\hat{\theta} = h(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_q) \simeq h(\tau_1, \tau_2, \ldots, \tau_q) + \sum_{j=1}^{q} (\hat{\tau}_j - \tau_j) \frac{\partial h}{\partial \tau_j}(\tau_1, \tau_2, \ldots, \tau_q).$$

Nótese que este es un estimador aproximadamente insesgado de θ, ya que $\hat{\tau}_j$ es un estimador insesgado de τ_j, y que de tomarse $a_j = \frac{\partial h}{\partial \tau_j}(\tau_1, \tau_2, \ldots, \tau_q)$, se cumple que

$$\hat{\theta} \simeq \theta + \sum_{j=1}^{q} a_j (\hat{\tau}_j - \tau_j).$$

Tomándose la varianza a la última expresión se tiene que

$$V(\hat{\theta}) \simeq \sum_{j=1}^{q} a_j^2 Var(\hat{\tau}_j) + 2 \sum_{j=1}^{q} \sum_{h=j+1}^{q} a_j a_h Cov(\hat{\tau}_j, \hat{\tau}_h). \tag{5.3}$$

Luego, una estimación de esta varianza puede obtenerse estimando los a_j y las varianzas y covarianzas de los estimadores de los totales.

5.4.2. El estimador de razón y regresión

Como ilustración de la técnica de linealización, pensemos en un estimador que ha sido recurrentemente utilizado a lo largo del texto. Este viene dado por el cociente o la razón de la estimación de totales o medias de dos variables x e y

$$\hat{\theta} = \frac{\hat{\tau}_y}{\hat{\tau}_x} = \frac{\bar{Y}}{\bar{X}}$$

y por ello recibe el nombre de estimador de razón. La fórmula (5.3) nos provee, entonces, de la siguiente aproximación para la varianza de este estimador:

$$V(\hat{\theta}) = \theta^2 \left(\frac{V(\hat{\tau}_x)}{\tau_x^2} + \frac{V(\hat{\tau}_y)}{\tau_y^2} - \frac{2Cov(\hat{\tau}_x, \hat{\tau}_y)}{\tau_x \tau_y} \right), \tag{5.4}$$

donde el parámetro $\theta = \frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}$ es el cociente de los totales o medias de las variables x e y en la población.

En muchos casos es común que el interés al utilizar un estimador como este se centre en alguna de sus variables; por decir, y, y que la otra variable x actúe como una variable auxiliar que si estuviera correlacionada con y y su total poblacional τ_x fuese conocido, nos podría ser de mucha utilidad para mejorar las estimaciones de la media o del total de y (mediante $\hat{\tau}_{ry} = \hat{\theta}\tau_x$) e incluso del mismo estimador de razón. El hecho que τ_x, o la media poblacional de la potencial variable predictora x, μ_x, se conozca puede parecer extraño; pero podría ocurrir (y sucede usualmente) que tal información esté consignada en el marco muestral o sea de fácil acceso. Puede también ocurrir que se disponga de esta información de un censo o estudio previo sobre la misma población. Explicitemos seguidamente la varianza de este estimador en el caso de un MASs de tamaño n de una población de tamaño N. Dado que $\mu_y = \theta\mu_x$, la ecuación (5.4) podrá reescribirse como

$$V(\hat{\theta}) = \theta^2 \left(\frac{V(\bar{X})}{\mu_x^2} + \frac{V(\bar{Y})}{\theta^2\mu_x^2} - \frac{2Cov(\bar{X},\bar{Y})}{\mu_x\theta\mu_x} \right)$$

$$= \frac{1}{\mu_x^2}\left(\theta^2(1-\frac{n}{N})\frac{\sigma_x^2}{n} + (1-\frac{n}{N})\frac{\sigma_y^2}{n} - 2\theta(1-\frac{n}{N})\frac{\sigma_{xy}}{n} \right) = \frac{1}{n\mu_x^2}(1-\frac{n}{N})\left(\sigma_y^2 + \theta^2\sigma_x^2 - 2\theta\sigma_{xy}\right)$$

Así, un estimador de esta varianza puede obtenerse mediante

$$\hat{V}(\hat{\theta}) = \frac{1}{n\mu_x^2}(1-\frac{n}{N})(S_y^2 + \hat{\theta}^2 S_x^2 - 2\hat{\theta}S_{xy}),$$

o, alternativamente, al definirse $\hat{z}_i = y_i - \hat{\theta}x_i$, mediante

$$\hat{V}(\hat{\theta}) = \frac{1}{\mu_x^2}(1-\frac{n}{N})\frac{\hat{\sigma}_z^2}{n}, \tag{5.5}$$

donde $\hat{\sigma}_z^2 = \frac{1}{n-1}\sum_{i=1}^{N}(\hat{z}_i - \hat{\bar{z}})^2\delta_i$ y $\hat{\bar{z}}$ es la media muestral de los $\hat{z}_i$. En ambas expresiones, si la media poblacional μ_x no se conociese, ella podría reemplazarse por $\bar{X}$.

Es interesante observar que podríamos haber también deducido la varianza de este estimador, si hubiéramos considerado que el sesgo del estimador puede aproximarse por

$$\hat{\theta} - \theta = \frac{\bar{Y} - \theta\bar{X}}{\bar{X}} \simeq \frac{\bar{Y} - \theta\bar{X}}{\mu_x}.$$

Luego, al tomársele la varianza a esta expresión se obtiene la aproximación $V(\hat{\theta}) = \frac{1}{\mu_x^2}V(\bar{Z}) = \frac{1}{\mu_x^2}(1-\frac{n}{N})\frac{\sigma_z^2}{n}$, donde $\bar{Z} = \bar{Y} - \theta\bar{X}$ y σ_z^2 denota la varianza de los $z_i = y_i - \theta x_i$ en la población.

En un MAE, el argumento anterior requiere de cierto cuidado. La mayoría de softwares y la literatura sugieren que se utilice un estimador de razón *combinado*, esto es, un estimador de la forma

$$\hat{\theta} = \frac{\bar{Y}}{\bar{X}} = \frac{\sum_{h=1}^{H}\frac{N_h}{N}\bar{Y}_h}{\sum_{h=1}^{H}\frac{N_h}{N}\bar{X}_h},$$

el cual difiere de uno *separado* en que los ratios por estrato se promedian ponderadamente a través de $\hat{\theta}_s = \sum_{h=1}^{H} \frac{N_h}{N}\hat{\theta}_h = \sum_{h=1}^{H} \frac{N_h}{N}\frac{\bar{Y}_h}{\bar{X}_h}$. Para hacer una comparación entre estos estimadores véase Cochran (1977). Nosotros, a falta de aclaración, utilizaremos siempre el primero. La varianza aproximada del estimador de razón combinado se puede obtener por un argumento similar al del MASs; esto es, tomándose la varianza a la siguiente aproximación del sesgo del estimador

$$\hat{\theta} - \theta = \frac{\bar{Y} - \theta\bar{X}}{\bar{X}} = \frac{\sum_{h=1}^{H} \frac{N_h}{N}(\bar{Y}_h - \theta\bar{X}_h)}{\bar{X}} \simeq \frac{\sum_{h=1}^{H} \frac{N_h}{N}(\bar{Y}_h - \theta\bar{X}_h)}{\mu_x}.$$

Así, uno obtiene que aproximadamente

$$V(\hat{\theta}) = \frac{1}{\mu_x^2}\sum_{h=1}^{H}(\frac{N_h}{N})^2(1 - \frac{n_h}{N_h})\frac{\sigma_{hz}^2}{n_h}, \tag{5.6}$$

siendo $\sigma_{hz}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(z_{ih} - \mu_{hz})^2$ la varianza de todos los $z_{ih} = y_{ih} - \theta x_{ih}$ en el estrato h.

Un estimador de esta última varianza viene dado por

$$\hat{V}(\hat{\theta}) = \frac{1}{\mu_x^2}\sum_{h=1}^{H}(\frac{N_h}{N})^2(1 - \frac{n_h}{N_h})\frac{\hat{\sigma}_{hz}^2}{n_h}, \tag{5.7}$$

siendo $\hat{\sigma}_{hz}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(z_{ih} - \mu_{hz})^2\delta_{hi}$ la varianza muestral de todos los $z_{ih} = y_{ih} - \hat{\theta}x_{ih}$, con $i = 1, 2 \ldots, n_h$, en el estrato h. Similarmente, de no conocerse μ_x, este podría reemplazarse por $\bar{X}$.

5.4.3. Métodos de remuestreo

Otro enfoque para la estimación de varianzas se basa en el uso de técnicas de remuestreo. La idea aquí es obtener varias estimaciones del parámetro de interés θ mediante replicación de partes comparables de la muestra original y usar la variabilidad de tales estimaciones para estimar la varianza del estimador θ.

Para tener una idea de cómo funciona ello consideremos m estimadores insesgados y no correlacionados $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ de θ. La media aritmética de estos estimadores

$$\bar{\theta} = \frac{1}{m}\sum_{r=1}^{m}\hat{\theta}_r$$

es claramente otro estimador insesgado de θ, y su varianza viene dada por

$$V(\bar{\theta}) = \frac{1}{m^2}\sum_{r=1}^{m}V(\hat{\theta}_r).$$

La siguiente proposición nos brinda un estimador insesgado de esta varianza.

Proposición 5.1. *Un estimador insesgado de $V(\bar{\theta})$ viene dado por*

$$\hat{V}(\bar{\theta}) = \frac{1}{m(m-1)} \sum_{r=1}^{m} (\hat{\theta}_r - \bar{\theta})^2.$$

Demostración: *Tomándose el valor esperado al estimador propuesto, se tiene que*

$$E(\hat{V}(\bar{\theta})) = \frac{1}{m(m-1)} \sum_{r=1}^{m} E((\hat{\theta}_r - \bar{\theta})^2).$$

Restando y sumando θ al interior del valor esperado a derecha resulta que

$$E((\hat{\theta}_r - \bar{\theta})^2) = E((\hat{\theta}_r - \theta)^2) + E((\bar{\theta} - \theta)^2) - 2E((\hat{\theta}_r - \theta)(\bar{\theta} - \theta))$$

$$= V(\hat{\theta}_r) + V(\bar{\theta}) - 2Cov(\hat{\theta}_r, \frac{1}{m} \sum_{r=1}^{m} \hat{\theta}_r) = V(\hat{\theta}_r) + \frac{1}{m^2} \sum_{r=1}^{m} V(\hat{\theta}_r) - \frac{2}{m} V(\hat{\theta}_r).$$

Así,

$$E(\hat{V}(\bar{\theta})) = \frac{1}{m(m-1)}((1 - \frac{2}{m}) \sum_{r=1}^{m} V(\hat{\theta}_r) + \frac{1}{m} \sum_{r=1}^{m} V(\hat{\theta}_r)) = \frac{1}{m^2} \sum_{r=1}^{m} V(\hat{\theta}_r) = V(\bar{\theta}). \quad \blacksquare$$

Exploraremos seguidamente en este capítulo tres de las técnicas de remuestreo más utilizadas: el muestreo por mitades balanceadas o BRR (de balanced repeated replication), el método Jacknife y el método Bootstrap.

5.4.4. El muestreo por mitades balanceado

Esta técnica, aplicable inicialmente a diseños estratificados donde se seleccionan al azar y con reemplazamiento a $n_h = 2$ unidades primarias (usualmente conglomerados) de las N_h en cada estrato h, fue propuesta por McCarthy (1969), quién se inspiró en los diseños multifactoriales propuestos por Plackett y Burman (1946). Si H denota el número de estratos y el interés radica en estimar la media poblacional μ, sabemos por lo estudiado en el capítulo 3 que el estimador insesgado de este viene dado por

$$\bar{Y}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_h,$$

donde $N = \sum_{h=1}^{H} N_h$, $\bar{Y}_h = \frac{1}{2}(Y_{h1} + Y_{h2})$ e Y_{h1}, Y_{h2} denotan el valor que tendrá y en las unidades seleccionadas del estrato h, las que, sin pérdida de generalidad, asumiremos que han sido seleccionadas secuencialmente. Recordemos también que el estimador natural de la varianza de este estimador viene dado por

$$\hat{V}(\bar{Y}_{st}) = \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{S_h^2}{2} = \frac{1}{4} \sum_{h=1}^{H} (\frac{N_h}{N})^2 D_h^2,$$

donde $D_h = Y_{h1} - Y_{h2}$.

Note que $\bar{Y}_{st}$ puede escribirse también como $\bar{Y}_{st} = \frac{1}{2}(\bar{Y}_{st,r} + \bar{Y}_{st,nr})$, donde $\bar{Y}_{st,r} = \sum_{h=1}^{H} \frac{N_h}{N} Y_{h1}$ e $\bar{Y}_{st,rc} = \sum_{h=1}^{H} \frac{N_h}{N} Y_{h2}$ son también estimadores insesgados e independientes de μ. En tal sentido, podríamos invocar la proposición 5.1 y obtener el siguiente estimador insesgado de $V(\bar{Y}_{st})$:

$$\hat{V}_r(\bar{Y}_{st}) = \frac{1}{2}((\bar{Y}_{st,r} - \bar{Y}_{st})^2 + (\bar{Y}_{st,rc} - \bar{Y}_{st})^2) = (\bar{Y}_{st,r} - \bar{Y}_{st})^2 = \frac{1}{4}(\bar{Y}_{st,r} - \bar{Y}_{st,rc})^2. \tag{5.8}$$

Como se ve, este estimador es más simple que $\hat{V}(\bar{Y}_{st})$, pero es menos eficiente. A fin de mejorar su eficiencia optaremos por considerar la metodología de replicación por mitades. La idea es generar réplicas al dividir la muestra tomada en dos mitades, las cuales estamos denotando por r y rc. Estas réplicas se construyen asignando una de las dos unidades primarias seleccionadas de cada estrato a la primera mitad y dejando la unidad primaria restante para la otra mitad. Observe que existen un total de 2^H asignaciones posibles o réplicas por mitades como esta. De modo resumido, el muestreo por mitades balanceados o BBR nos brindará un estimador de $V(\bar{Y}_{st})$ resultante de promediar los estimadores (5.8) para todas las distintas réplicas (o como más delante veremos, para un subconjunto apropiado de ellas). Este estimador viene dado por

$$\hat{V}_{BRR}(\bar{Y}_{st}) = \frac{1}{2^H} \sum_{r=1}^{2^H} \hat{V}_r(\bar{Y}_{st}) = \frac{1}{2^H} \sum_{r=1}^{2^H} (\bar{Y}_{st,r} - \bar{Y}_{st})^2. \tag{5.9}$$

Ejemplo 5.4. *Para una mejor comprensión consideremos el siguiente ejemplo de un MAE con 4 estratos, en el que se han observado los siguientes resultados:*

Estrato (h)	*Tamaño del estrato (N_h)*	y_{h1}	y_{h2}	$\bar{y}_h$	$d_h = y_{h1} - y_{h2}$
1	*300*	*235*	*179*	*185*	*56*
2	*100*	*525*	*483*	*504*	*42*
3	*50*	*950*	*1350*	*1150*	*- 400*
4	*200*	*759*	*990*	*875*	*-231*

Trabajando con una precisión de tres decimales, la media estimada de la población resulta ser $\bar{y}_{st}$ =530.615; mientras que la estimación de la varianza $V(\bar{Y}_{st})$ por MAE resulta ser 1677.112. Este será nuestro valor de referencia. Lo primero a notar es que la estimación (5.8) nos da un valor de 1208.899 que es distinto a 1677.112. El número de réplicas por mitades para este problema es $2^4 = 16$, siendo una réplica distinta a la anterior, por ejemplo, $\{y_{11}, y_{21}, y_{32}, y_{42}\}$ para la primera mitad, que aquí la denotaremos por A. A la otra mitad la denotaremos por B. Con esta réplica, la estimación (5.8) resulta ser 4499.314. Como se aprecia, hay bastante diferencia en esta estimación con la de la réplica anterior y son, precisamente, estas distintas estimaciones las que nos permitirán obtener una mejor estimación de $V(\bar{Y}_{st})$ al promediarlas como en (5.9) sobre todas las réplicas. La figura 5.1 muestra

Fila			h	Nh	yh1	yh2	ybarh	dh	Nh/N								
1			1	300	235	179	207	56	0.462								
2			2	100	525	483	504	42	0.154		Estimador clásico de la media =					530.615	
3			3	50	950	1,350	1,150	-400	0.077								
4			5	200	759	990	875	-231	0.308								
5	Estimador			650	495.846	565.385											
6	de la varianza MAE				1677.112												
7	de la varianza con una réplica				1208.899												
8																	
9										Réplicas							
10	Estratos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
11	1	1	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1
12	2	1	1	1	-1	1	1	-1	-1	-1	1	1	-1	1	-1	-1	-1
13	3	1	1	-1	1	1	-1	-1	1	1	1	-1	-1	-1	1	-1	-1
14	4	1	-1	1	1	1	-1	1	1	-1	-1	1	-1	-1	-1	1	-1
15	Estratos																
16	1	235	235	235	235	179	235	235	179	235	179	179	235	179	179	179	179
17	2	525	525	525	483	525	525	483	483	483	525	525	483	525	483	483	483
18	3	950	950	1350	950	950	1350	1350	950	950	950	1350	1350	1350	950	1350	1350
19	4	759	990	759	759	759	990	759	759	990	990	759	990	990	990	759	990
20																	
21	Media réplica mitad																
22	A	495.846	566.923	526.615	489.385	470.000	597.692	520.154	463.538	560.462	541.077	500.769	591.231	571.846	534.615	494.308	565.385
23	B	565.385	494.308	534.615	571.846	591.231	463.538	541.077	597.692	500.769	520.154	560.462	470.000	489.385	526.615	566.923	495.846
24	vr	1208.899	1318.249	16.000	1699.976	3674.225	4499.314	109.444	4499.314	890.793	109.444	890.793	3674.225	1699.976	16.000	1318.249	1208.899
25					530.615												
26				varBRR =	1677.112												

Figura 5.1: Muestreo por mitades balanceado para cuatro estratos

las 16 réplicas por mitades existentes para este problema, donde en la fila 24 se tiene la estimación (5.8) para cada réplica. En la fila 25 se aprecia que el promedio de las medias para las réplicas de la mitad A coincide con la estimación por MAE de la media y, lo más sorprendente, el promedio (5.9) de las varianzas para las 16 réplicas es exactamente igual al valor de referencia en la estimación por el MAE. □

Con el fin de generalizar los resultados del ejemplo anterior, introduzcamos para cada réplica r la variable auxiliar δ_{hr}, que toma el valor 1 si la unidad Y_{h1} del estrato h está en la primera mitad de esta réplica. En caso contrario, δ_{hr} valdrá 0. Así, el estimador de la media poblacional para la r-ésima réplica en su primera mitad viene dado por

$$\bar{Y}_{st,r} = \sum_{h=1}^{H} \frac{N_h}{N}(Y_{h1}\delta_{hr} + Y_{h2}(1-\delta_{hr})).$$

Definamos ahora la variable aleatoria

$$\delta_h^{(r)} = 2\delta_{hr} - 1 = \begin{cases} 1 & \text{si } Y_{h1} \text{ está en la mitad A de la réplica r} \\ -1 & \text{si } Y_{h2} \text{ está en la mitad A de la réplica r} \end{cases}$$

Note que estas variables satisfacen por construcción que $\sum_{r=1}^{2^H} \delta_h^{(r)} = 0$ y

$$\sum_{r=1}^{2^H} \delta_h^{(r)} \delta_l^{(r)} = 0 \tag{5.10}$$

para cualesquiera de los estratos $h \neq l$ en la población. Más aún, se cumple que

$$\bar{Y}_{st,r} - \bar{Y}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \delta_h^{(r)} \frac{D_h}{2}.$$

Estamos ya entonces en condiciones de establecer la siguiente proposición que formaliza lo visto en nuestro ejemplo.

Proposición 5.2. *En un MAE con reemplazamiento de dos unidades seleccionadas por estrato se cumple que*

a)

$$\frac{1}{2^H} \sum_{r=1}^{2^H} \bar{Y}_{st,r} = \bar{Y}_{st}$$

b)

$$\hat{V}_{BRR}(\bar{Y}_{st}) = \hat{V}(\bar{Y}_{st})$$

Demostración: *Puesto que*

$$\sum_{r=1}^{2^H} \delta_{hr} = 2^{H-1},$$

se tiene que

$$\frac{1}{2^H}\sum_{r=1}^{2^H} \bar{Y}_{st,r} = \frac{1}{2^H}\sum_{h=1}^{H} \frac{N_h}{N}(Y_{h1}(\sum_{r=1}^{2^H} \delta_{hr}) + Y_{h2}(2^H - \sum_{r=1}^{2^H} \delta_{hr})) = \sum_{h=1}^{H}(\frac{Y_{h1}+Y_{h2}}{2}) = \bar{Y}_{st}.$$

Más aún,

$$(\bar{Y}_{st,r} - \bar{Y}_{st})^2 = \sum_{h=1}^{H}(\frac{N_h}{N})^2\frac{D_h^2}{4} + \sum_{\substack{h=1 \\ l\neq h}}^{H}\sum_{l=1}^{H} \frac{N_h}{N}\frac{N_l}{N}\delta_h^{(r)}\delta_l^{(r)}\frac{D_h D_l}{2}$$

y, por tanto, podemos escribir (5.9) como

$$\hat{V}_{BRR}(\bar{Y}_{st}) = \frac{1}{2^H}\sum_{r=1}^{2^H} \hat{V}_r(\bar{Y}_{st}) = \frac{1}{2^H}\sum_{r=1}^{2^H}(\bar{Y}_{st,r} - \bar{Y}_{st})^2$$

$$= \sum_{h=1}^{H}(\frac{N_h}{N})^2\frac{D_h^2}{4} + \frac{1}{2^H}\sum_{\substack{h=1 \\ l\neq h}}^{H}\sum_{l=1}^{H} \frac{N_h}{N}\frac{N_l}{N}\frac{D_h D_l}{2}(\sum_{r=1}^{2^H} \delta_h^{(r)}\delta_l^{(r)})$$

Consecuentemente, una aplicación directa de (5.10) nos conduce a que

$$\hat{V}_{BRR}(\bar{Y}_{st}) = \frac{1}{4}\sum_{h=1}^{H}(\frac{N_h}{N})^2 D_h^2 = \hat{V}(\bar{Y}_{st}).$$ ■

Claramente, una desventaja del estimador BRR es que cuando H es grande, este resulta inmanejable. Afortunadamente, es posible mostrar que para algunos valores de H, en concreto para valores enteros múltiplos de 4, una selección adecuada de tan solo $k = H$ de estas réplicas nos permitirá obtener exactamente el mismo estimador $\hat{V}(\bar{Y}_{st})$ que si consideramos todas las 2^H réplicas. Como la ecuación (5.10) y la demostración de la proposición anterior lo sugieren, para este subconjunto de k réplicas se deberá cumplir que

$$\sum_{r=1}^{k} \delta_h^{(r)}\delta_l^{(r)} = 0,$$

cualesquiera sean los estratos $h \neq l$ en la población. En tal caso se dice que las réplicas están en balance ortogonal, ya que la matriz cuadrada de orden $H \times H$, $[\delta_{hr}]$, llamada también matrix de Hadamard, es ortogonal. Estas matrices se encuentran tabuladas en distintos textos y se conjetura que existen para todo orden múltiplo de 4; siendo la de orden 200 la mayor hasta el momento construida.

Ejemplo 5.5. *Mostraremos para el ejemplo 5.4. un balance ortogonal con $H = 4$ estratos. Este y su estimación se muestran a continuación:*

	Réplicas				*Réplicas*			
Estratos	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*
1	*1*	*1*	*1*	*1*	*235*	*235*	*235*	*235*
2	*1*	*-1*	*1*	*-1*	*525*	*483*	*525*	*483*
3	*1*	*1*	*-1*	*-1*	*950*	*950*	*1350*	*1350*
4	*1*	*-1*	*-1*	*1*	*759*	*990*	*990*	*759*
	Media primera mitad				*495.846*	*560.462*	*597.692*	*520.154*
	d_h^2				*1208.899*	*890.793*	*4499.314*	*109.444*
	Varianza estimada BRR = 1677.112							

Como se aprecia, la varianza estimada de la media bajo estas 4 réplicas coincide con la estimación de referencia del MAE. La implementación en R del método BRR para este ejemplo viene dada por

```
mR = data.frame(y = c(235,525,950,759,179,483,1350,990),
Estrato = rep(1:4,2),Nh = rep(c(300,100,50,200),2))
mR$w = mR$Nh/2
(dism = svydesign(ids=~1, strata=~Estrato,weights =~w,data=mR))

## Stratified Independent Sampling design (with replacement)
## svydesign(ids = ~1, strata = ~Estrato, weights = ~w, data = mR)

# Convirtiendo el diseño para remuestreo
(dBRR = as.svrepdesign(design=dism,type="BRR"))

## Call: as.svrepdesign(design = dism, type = "BRR")
## Balanced Repeated Replicates with 8 replicates.

(mm = svymean(~y,design=dBRR))

##   mean SE
## y  531 41

# Varianza buscada
SE(mm)^2

## [1] 1677
```

□

Observaciones:

- En la práctica, el número de estratos H no necesariamente es múltiplo de 4, por lo cual por lo que la existencia de una matriz de Hadamard no está garantizada. Afortunadamente, se pueden implementar sobre la base de los diseños de Plackett y Burman (1946) y, tal como se hace en R, algoritmos que generan un número k de réplicas igual al menor múltiplo de 4 que sea mayor que H, generándose con ellas una matriz de pesos de réplica de orden $H \times k$, cuyas columnas no necesariamente son ortogonales, pero satisfacen aun la propiedad de brindar la correcta estimación de la varianza.

- El tamaño de muestra de unidades primarias por estrato no necesita ser exactamente $n_h = 2$. Si este fuera el caso, se podría forzar la situación anterior segmentando, por ejemplo, el estrato h en estratos artificiales de, aproximadamente, igual tamaño y tomándose luego al azar y con reemplazamiento 2 de estos pseudoestratos.

- Si bien el método BRR nos brinda una estimación exacta en la estimación de la varianza de estimadores como la media o el total, esto solo se cumplirá aproximadamente para otros estimadores no lineales $\hat{\theta}$. El cómputo del estimador para cada réplica r se hace en la práctica con los pesos de réplica, los cuales ajustan a los pesos de muestreo ω. El ajuste para toda unidad i seleccionada en el estrato h se hace mediante

$$\omega_{hi}(r) = \begin{cases} 2\omega_{hi} & \text{si la unidad } i \text{ está en la primera mitad de la réplica r} \\ 0 & \text{en caso contrario} \end{cases}$$

  desde que existe igual probabilidad de que la unidad i sea asignada o no a la primera mitad. Estos pesos se usan luego para construir la correspondiente función de probabilidad empírica y el estimador $\hat{\theta}_{(r)}$ que tiene la misma forma que $\hat{\theta}$, pero con pesos distintos. El estimador de varianza BRR para la varianza de $\hat{\theta}$ viene, similarmente a (5.9), dado por

$$\hat{V}_{BRR}(\hat{\theta}) = \frac{1}{k}\sum_{r=1}^{k}(\hat{\theta}_{(r)} - \hat{\theta})^2. \tag{5.11}$$

Cabe comentar que el código del ejemplo 5.6 ha usado el comando `as.svrepdesign` a fin de convertir el diseño original en uno de remuestreo. Alternativamente, uno podría definir de forma directa el diseño de remuestreo con el comando `svrepdesign`. Ello es útil cuando la base de datos incluye como información los pesos de réplica y los pesos de muestreo. Los pesos de muestreo se usan para el cálculo del estimador puntual y los de réplica para el de su varianza. Los pesos de réplicas en el ejemplo 5.6 pueden obtenerse mediante

```
(Wr = weights(dBRR))

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    2    0    2    0    2    0    2    0
## [2,]    2    2    0    0    2    2    0    0
## [3,]    2    0    0    2    2    0    0    2
## [4,]    2    2    2    2    0    0    0    0
## [5,]    0    2    0    2    0    2    0    2
## [6,]    0    0    2    2    0    0    2    2
## [7,]    0    2    2    0    0    2    2    0
## [8,]    0    0    0    0    2    2    2    2
```

Note aquí que solo se muestran los pesos de réplica sin el ajuste a los pesos. Esta matriz es siempre de orden $2H \times k$, pues contiene en las columnas las réplicas ortogonales o generadas por R (véase la primera observación); y, en las filas, las unidades consideradas para ambas mitades. Si deseamos utilizar el comando `svrepdesign` para obtener los mismo resultados que en el ejemplo 5.6, podríamos escribir indistintamente cualquiera de las siguientes líneas:

```
(dBRRa<-svrepdesign(data=mR,type="BRR",repweights=Wr,weights=~w,
combined.weights=FALSE))

## Call: svrepdesign.default(data = mR, type = "BRR", repweights = Wr,
##     weights = ~w, combined.weights = FALSE)
## Balanced Repeated Replicates with 8 replicates.

(dBRRb<-svrepdesign(data=mR, type="BRR", repweights=Wr*mR$w,weights=~w))

## Call: svrepdesign.default(data = mR, type = "BRR", repweights = Wr *
##     mR$w, weights = ~w)
## Balanced Repeated Replicates with 8 replicates.
```

donde en el segundo caso los pesos se dan en su forma ajustada o combinada $\omega_{hi}(r)$. En efecto, se cumple que

```
svymean(~y,dBRRa)

##   mean SE
## y  531 41

svymean(~y,dBRRb)
```

```
##   mean SE
## y  531 41
```

brindan las mismas estimaciones que las obtenidas en el ejemplo 5.6. Estos resultados podrían también obtenerse sin usar el paquete survey. Si empleamos las mismas réplicas aquí utilizadas, el código correspondiente sería

```
mRB = cbind(mR,Wr) # Base de datos con pesos de replica
mer = 0
for(i in 1:8){
 r = by(mRB$w*mRB[,4+i],mRB$y,sum)
 Phat = as.vector(r/sum(r))
 mer[i] = sum(as.numeric(names(r))*Phat)}
c(mean(mer),sqrt(mean((mer-mean(mer))^2)))

## [1] 531  41
```

Una limitación de la metodología BRR es que una de las muestras por mitades es siempre eliminada al formar una réplica. Ello podría ocasionar inestabilidad en la estimación de la varianza del estimador en el caso de que se consideren, por ejemplo, dominios de estudio, pues podría ocurrir que todo el dominio ocurra precisamente en las mitad eliminada de una réplica particular. Para evitar situaciones como esta, Fay (1984) y Dippo et al. (1984) propusieron modificar el método incluyendo todas las observaciones en cada réplica mediante la asignación de pesos de réplica $\omega_{hi}(r) = (2-\rho)\omega_{hi}$, si la unidad i del estrato h es seleccionada en la primera mitad de la réplica r, y pesos $\omega_{hi}(r) = \rho\omega_{hi}$, en caso contrario. Aquí $\rho \in [0, 1[$ es un parámetro por fijar, siendo $\rho = 0.3$ una elección común.

Diversas extensiones de la metodología BRR puede consultarse en Wolter (2007). Allí se estudia, por ejemplo, cómo modificar el remuestreo si la selección se hace sin reemplazamiento, cómo seleccionar más de dos unidades primarias por estrato y cómo adaptar estos procedimientos a diversos esquemas de muestreo complejo.

5.4.5. El método Jackknife

Esta técnica, introducida inicialmente por Quenouille (1949) para la reducción de sesgo en series temporales y desarrollada posteriormente por Tukey (1958), consiste en particionar la muestra de tamaño n en J grupos y estimar igual cantidad de veces el parámetro de interés θ después de haberse eliminado o cortado en cada ocasión a uno de los grupos. La variabilidad entre estas estimaciones pueden entonces usarse luego para estimar la variabilidad del estimador original propuesto para θ.

Como ejemplo, consideremos un diseño multietápico estratificado con H estratos y en los que se seleccionen con reemplazamiento n_h UPM de cada estrato h. Si θ es el parámetro de interés y $\hat{\theta}$ su estimador basado en la totalidad de la muestra, denotemos por $\hat{\theta}_{(hj)}$ al estimador de θ basado en la muestra luego de omitir a (toda) la UPM j del estrato h. El estimador Jackknife de la varianza de $\hat{\theta}$ viene dado por

$$\hat{V}_{JKn}(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\hat{\theta}_{(hj)} - \hat{\theta})^2. \tag{5.12}$$

En la práctica, este estimador se calcula con los pesos de réplica. Si se remueve la UPM j para formar la réplica (hj) en el estrato h y ω_i^0 es el peso (base) de una unidad no primaria i, entonces los pesos de réplicas ajustan estos mediante

$$\omega_{i(hj)} = \begin{cases} 0 & \text{si } i \text{ está en la UPM } j \text{ del estrato h} \\ \frac{n_h}{n_h - 1}\omega_i^0 & \text{si } i \text{ está en el estrato } h \text{ pero no en la UPM } j \\ \omega_i^0 & \text{si } i \text{ no está en el estrato } h \end{cases}$$

Estos pesos ajustados se emplean luego para construir la correspondiente función de probabilidad empírica y el cálculo de $\hat{\theta}_{(hj)}$. Finalmente, estos pesos se reemplazan en (5.12) para calcular la estimación de la varianza.

Cabe comentar un caso particular del estimador Jacknife al que se suele denotar en R por JKn; este es el no estratificado JK1 que se obtiene cuando $H = 1$. Para este, el estimador toma la forma

$$\hat{V}_{JK1}(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^{n} (\hat{\theta}_{(j)} - \hat{\theta})^2.$$

En el caso de la estimación de la media $\theta = \mu$ mediante un MASc, se tiene que para $\hat{\theta} = \bar{Y}$ se cumple que $\hat{\theta}_{(j)} = \frac{1}{n-1} \sum_{i \neq j}^{n} Y_i = \bar{Y} - \frac{1}{n-1}(Y_j - \bar{Y})$. Así,

$$\hat{V}_{JK1}(\bar{Y}) = \frac{n-1}{n} \sum_{i=1}^{n} (\bar{Y} - \frac{1}{n-1}(Y_j - \bar{Y}) - \bar{Y})^2 = \frac{S^2}{n},$$

el cual es, por las proposiciones 2.1 y 2.2, el estimador natural insesgado de la varianza de $\bar{Y}$ en un MASc, y de allí la inclusión del término $\frac{n-1}{n}$ en el estimador.

Ejemplo 5.6. *Para ilustrar esta técnica y compararla con la anterior retomemos el MAE del ejemplo 5.5 para el cual creamos en el ejemplo 5.6 el diseño* `dism`. *Este diseño clásico se podrá convertir en uno de remuestreo Jackknife y nos calculará el estimador (5.12) para la media mediante*

```
(dJKn = as.svrepdesign(design=dism,type="JKn"))

## Call: as.svrepdesign(design = dism, type = "JKn")
## Stratified cluster jackknife (JKn) with 8 replicates.

(mm = svymean(~y,design=dJKn))

##   mean SE
## y  531 41

# Estimacion (5.12)
SE(mm)^2

## [1] 1677
```

Al igual que con el método BRR, dos maneras alternativas de obtener esta estimación serán con el comando `svrepdesign` *o programándola directamente en R a través del desarrollo anterior. Los códigos son:*

```
# Pesos no ajustados de replicas con Jackknife
(Wr = weights(dJKn))

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    0    2    1    1    1    1    1    1
## [2,]    1    1    0    2    1    1    1    1
## [3,]    1    1    1    1    0    2    1    1
## [4,]    1    1    1    1    1    1    0    2
## [5,]    2    0    1    1    1    1    1    1
## [6,]    1    1    2    0    1    1    1    1
## [7,]    1    1    1    1    2    0    1    1
## [8,]    1    1    1    1    1    1    2    0

(dJKna<-svrepdesign(data=mR,type="JKn",repweights=Wr,weights=~w,scale=1,
rscales=0.5,combined.weights=FALSE))

## Call: svrepdesign.default(data = mR, type = "JKn", repweights = Wr,
##     weights = ~w, scale = 1, rscales = 0.5, combined.weights = FALSE)
## Stratified cluster jackknife (JKn) with 8 replicates.

(dJKnb<-svrepdesign(data=mR, type="JKn",repweights=Wr*mR$w,weights=~w,
scale=1,rscales=0.5))
```

```
## Call: svrepdesign.default(data = mR, type = "JKn", repweights = Wr *
##     mR$w, weights = ~w, scale = 1, rscales = 0.5)
## Stratified cluster jackknife (JKn) with 8 replicates.

svymean(~y,dJKna)

##   mean SE
## y  531 41

svymean(~y,dJKnb)

##   mean SE
## y  531 41

mRJ = cbind(mR,Wr) # Base con los pesos de replica
mer = 0
for(i in 1:8){
 r = by(mRJ$w*mRJ[,4+i],mRJ$y,sum)
 Phat = as.vector(r/sum(r))
 mer[i] = sum(as.numeric(names(r))*Phat)}
c(mean(mer),sqrt(sum((mer-mean(mer))^2)/2))

## [1] 531  41
```

□

5.4.6. El método Bootstrap

Esta es una técnica de remuestreo cuya lógica subyace en pensar la muestra como una población de la cual se extraen un gran número de submuestras bajo reemplazamiento, a las que llamamos réplicas. Estas réplicas finalmente se usan para estimar la varianza del estimador.

Existen distintas variantes del método Bootstrap en poblaciones finitas; pero solo algunas como la de Rao y Wu (1988), que aquí discutiremos, están implementadas en un software estadístico. Esta variante se aplica en la estimación de un parámetro θ mediante un estimador $\hat{\theta}$, no necesariamente lineal, bajo una muestra por conglomerados estratificada. Se recomienda usar entre $R = 500$ y $R = 1000$ réplicas y el método sigue los siguientes pasos:

1. Para cada estrato, seleccionar las R réplicas mediante un MASc de $n_h - 1$ UPM a partir de la muestra inicial de tamaño n_h de cada estrato h. Sea $m_{hj}(r)$ el número de veces que la j-ésima UPM del estrato h es seleccionado en la réplica r.

2. Para cada réplica $r = 1, 2, \ldots, R$ y para cada unidad i tomada de la UPM j del estrato h, reajustar los pesos como

$$\omega_{hji}(r) = \omega^0_{hji} \times \frac{n_h}{n_h - 1} m_{hj}(r),$$

donde ω^0_{hji} es el peso base para la unidad i perteneciente a la UPM j del estrato h.

3. Calcular el estimador $\hat{\theta}^*_r$ para la r-ésima réplica usando los pesos $\omega_{hji}(r)$.

4. El estimador de varianza bootstrap viene dado por

$$\hat{V}_B(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^{R} (\hat{\theta}^*_r - \hat{\theta})^2.$$

Una de las ventajas del método recae en su propiedad de generar aproximadamente la distribución de $\hat{\theta}$, lo cual nos permitirá determinar intervalos de confianza en forma directa. Para obtener un intervalo de confianza al 95 % podríamos, por ejemplo, considerar tan solo los percentiles 2.5 y 97.5 a partir de $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_R$. Otras técnicas alternativas para obtener estos intervalos pueden encontrarse en Efron y Tibshirani (1993).

Ejemplo 5.7. *Con el propósito de ilustrar y comparar los diferentes métodos de estimación de la varianza mostrados consideraremos el problema de la estimación de la varianza del rendimiento medio en Matemáticas para la DRE Amazonas en la ECE 2019 bajo un muestreo aleatorio por conglomerados estratificado. Para tal efecto, consideremos los mismos estratos que en el ejemplo 5.1 y seleccionemos dos conglomerados (colegios) por estrato. Seguidamente se muestran los códigos respectivos.*

```
Pop = ece19Am
Pop$Estrato=interaction(Pop$area,Pop$gestion2)
Pop = Pop[order(Pop$Estrato),]
set.seed(12345)
m=mstage(Pop,stage=list("stratified","cluster"),varnames=list("Estrato","ID_IE")
,size=list(size1=table(Pop$Estrato),size2 =c(2,2,2,2)),method=list("","srswor"))
mues = getdata(Pop,m)[[2]]
mues$w0 = 1/mues$Prob
```

Comparemos ahora la estimación del rendimiento medio en Matemáticas en Amazonas y obtengamos las estimaciones de los errores estándar de estimación bajo los cuatro métodos desarrollados.

```
dis19 = svydesign(id=~ID_IE, strata=~Estrato, nest=T,data=mues, probs=~Prob)
#Estimación por el método de linealización
r1 = svymean(~M500_CT,design=dis19,na.rm=T)
# Estimación BRR
brr19 = as.svrepdesign(design=dis19,type="BRR")
r2 = svymean(~M500_CT,design=brr19,na.rm=T)
#Estimación Jacknife
jkn19 = as.svrepdesign(design=dis19,type="JKn")
r3 = svymean(~M500_CT,design=jkn19,na.rm=T)
#Estimación Bootstrap
boot19 = as.svrepdesign(design=dis19,type="subbootstrap",replicates=1000)
r4 = svymean(~M500_CT,design=boot19,na.rm=T)
list(r1,r2,r3,r4)

## [[1]]
##         mean   SE
## M500_CT  520 18.2
##
## [[2]]
##         mean   SE
## M500_CT  520 18.2
##
## [[3]]
##         mean   SE
## M500_CT  520 18.3
##
## [[4]]
##         mean   SE
## M500_CT  520 18.6
```

□

5.5. Una introducción al análisis estadístico con muestras complejas

Hasta el momento hemos estudiado algunos estimadores puntuales de una variable. En una encuesta, sin embargo, uno no solo está interesado en cuestiones univariadas, sino en estudiar las distintas relaciones que se pudieran dar entre las variables incluidas en la en-

cuesta. En esta sección exploraremos tres de las áreas de mayor relevancia en el estudio de estas relaciones: el análisis de datos categóricos, el análisis de regresión y la comparación de una o más poblaciones

5.5.1. Análisis de datos categóricos con muestras complejas

Ya vimos que una distribución fundamental en el análisis de datos categóricos (es decir, de variables que solo pueden medirse en escala nominal u ordinal, como género, religión, ansiedad, nivel socioeconómico, etc.) es la distribución multinomial. Hipótesis sobre los parámetros de esta distribución se pueden traducir en distintos procedimientos estadísticos como las pruebas de independencia en tablas de contingencia, la igualdad de proporciones o las pruebas de bondad de ajuste.

Si $(X_1, X_2, \ldots, X_k) \sim Mul(n, p_1, p_2, \ldots, p_k)$, la prueba asintótica estándar para contrastar a nivel α

$$H_0 : p_1 = p_1^0,\ p_2 = p_2^0, \ldots, p_k = p_k^0 \ \text{ vs } \ H_1 : \exists i \ / \ p_i \neq p_i^0$$

donde los valores $p_1^0, \ldots, p_k^0$ son conocidos, es la prueba chi-cuadrado, cuya región crítica o de rechazo para H_0 viene dada por

$$\text{RC:} \quad U_0 = \sum_{i=1}^{k} \frac{(x_i - E_i^0)^2}{E_i^0} > \chi^2_{1-\alpha}(k-1),$$

donde x_i es el valor observado de X_i y $E_i^0 = np_i^0$ su frecuencia esperada bajo H_0. En muchas situaciones, sin embargo, las probabilidades p_i no son directamente especificadas en la hipótesis nula y dependen de otros parámetros poblacionales que requieren estimarse. Uno de estos casos se da en los contrastes sobre la independencia de dos variables categóricas X e Y, cuyos valores se encuentran tabulados en un tabla de contingencia. Pensemos, para contextualizar, que se ha tomado una encuesta por MASs a 500 de un total de 5000 hogares de cierto municipio para averiguar si la opinión acerca de la labor del alcalde distrital (variable Y) está asociada o no a que la familia posea o no vehículo particular (variable X). Tomada la encuesta al jefe de hogar, supongamos que sus respuestas se hayan resumido en la siguiente tabla de contingencia:

		Y		
	Auto particular	1 = Opinión desfavorable	2 = Opinión favorable	Total
X	1 = No	$n_{11} = 105$	$n_{12} = 188$	$n_{1.} = 207$
	2 = Sí	$n_{21} = 88$	$n_{22} = 119$	$n_{2.} = 293$
Total		$n_{.1} = 193$	$n_{.2} = 307$	$n = 500$

Las pruebas asintóticas más populares para contratar a nivel α

$$H_0 : X \text{ e } Y \text{ son independientes vs. } H_1 : X \text{ e } Y \text{ no son independientes}$$

son la prueba chi-cuadrado y la prueba de razón de verosimilitud. Ambas son asintóticamente equivalentes y sus regiones críticas para, en general, una tabla de contingencia con a categorías de X y b categorías de Y vienen dadas por

$$\text{R.C: } \chi_0^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(n_{ij}-\hat{E}_{ij}^0)^2}{\hat{E}_{ij}^0} > \chi_{1-\alpha}^2((a-1)(b-1))$$

y

$$\text{R.C: } G_0 = 2\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\log(\frac{n_{ij}}{\hat{E}_{ij}^0}) > \chi_{1-\alpha}^2((a-1)(b-1))$$

donde:

$$\hat{E}_{ij}^0 = n\hat{p}_{i.}^0\hat{p}_{.j}^0 = \frac{n_{i.}n_{.j}}{n}$$

es la estimación máxima verosímil de la frecuencia esperada en la celda (i, j) bajo H_0.

En nuestro ejemplo, los estadísticos de prueba correspondientes observados son χ_0^2 =2.281 y $G_0 = 2.275$; mientras que el valor en tabla de la distribución chi-cuadrado para $\alpha = 0.05$ es $\chi_{0.95}^2(1) = 3.84$. El valor p de este contraste es, por tanto, 0.131. Consecuentemente, no encontramos evidencia, en el municipio, de que la opinión hacia el alcalde tenga relación con el hecho de que la familia tenga o no auto particular. Estos análisis y las correspondientes salidas en R se muestran seguidamente:

```
Auto <-c(rep('No',193),rep('Si',307))
Opinion <- c(rep("Desfavorable",105),rep("Favorable",88),
rep("Desfavorable",188),rep("Favorable",119))
tt = table(Auto,Opinion)
summary(tt)

## Number of cases in table: 500
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 2.3, df = 1, p-value = 0.1
```

Una manera equivalente de plantear contrastes de independencia es mediante los odds ratios. En una tabla de contingencia 2×2 el odds ratio se define como

$$\theta = \frac{P(Y = 1 \mid X = 1)}{P(Y = 1 \mid X = 2)}.$$

Este parámetro poblacional puede tomar valores menores, iguales o mayores que 1. Un θ igual a 1 indica que no hay asociación y, por tanto, el contraste de independencia anterior equivale a plantear

$$H_0 : \theta = 1$$

contra una alternativa que incluso puede ser unilateral. Para una tabla de contingencia general $a \times b$, podríamos definir el vector columna $\boldsymbol{\theta} = [\theta_{11}, \theta_{12}, \ldots, \theta_{(a-1)(b-1)}]^\top$ con

$$\theta_{ij} = P(X = i, Y = j) - P(X = i)P(Y = j) \equiv p_{ij} - p_{i.}p_{.j}$$

y escribir la hipótesis de independencia como

$$H_0 : \boldsymbol{\theta} = 0.$$

Todo lo expuesto hasta el momento es válido bajo un MAS. Si el muestreo es complejo, las distribuciones de χ^2_0 y G_0 no serán más chi-cuadrado, lo cual nos podría llevar a conclusiones erróneas. En particular, la conglomeración tiene un fuerte efecto sobre estas distribuciones nulas. Para ilustrarlo retomemos el ejemplo previo, pero en el que no solo hayamos preguntado la opinión al jefe del hogar sino también a su cónyugue (note que ambos pertenecen a un mismo conglomerado, que es el hogar) y supongamos, exagerando (aunque la verdad no tanto), que ambos comparten la misma opinión sobre el alcalde; más explícitamente, que contamos con la siguiente tabla de contingencia:

		Y		
	Auto particular	1 = Opinión desfavorable	2 = Opinión favorable	Total
X	1 = No	$n_{11} = 210$	$n_{12} = 376$	$n_{1.} = 414$
	2 = Sí	$n_{21} = 176$	$n_{22} = 238$	$n_{2.} = 586$
Total		$n_{.1} = 386$	$n_{.2} = 614$	$n = 1000$

Note que se tiene aquí una correlación intraclase de 1.

Si evaluamos en este nuevo contexto nuestros estadísticos, obtendremos $\chi^2_0 = 4.562$ y $G_0 = 4.55$ que duplican a sus valores anteriores. Más aún, obtendremos un valor p de 0.03269 y, por tanto, podríamos estar tentados a concluir, equívocamente, de que sí existe asociación entre la opinión sobre el alcalde y la tenencia de auto particular. Note también que esto no es un fenómeno particular de este ejemplo, ya que, en general, bajo una correlación intraclase de 1 (por la duplicidad de respuestas en las unidades primarias) los estadísticos

$$\chi^2_0 = n \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(p_{ij} - \hat{p}^0_{i.}\hat{p}^0_{.j})^2}{\hat{p}^0_{i.}\hat{p}^0_{.j}}$$

y

$$G_0 = 2n \sum_{i=1}^{a} \sum_{j=1}^{b} p_{ij} \log(\frac{p_{ij}}{\hat{p}^0_{i.}\hat{p}^0_{.j}}),$$

siendo p_{ij} la proporción observada de respuestas en la celda (i, j), duplican su valor.

Veamos ahora dos procedimientos para incorporar el diseño en la prueba de independencia. Para ser más breves, nos centraremos en la prueba chi-cuadrado de Pearson, procedimientos similares existen para la prueba de razón de verosimilitud.

Para empezar, sea $(X_{11}, X_{12}, \ldots, X_{ab})$ el vector aleatorio en el que cada X_{ij} denota el número de elementos en la muestra de tamaño n que toman valores en la celda (i, j), siendo p_{ij} la probabilidad de que cualquier elemento de la muestra tome valores en esta celda. Sea, por otro lado, $\mathbf{p} = [p_{11}, p_{22}, \ldots, p_{ab-1}]^\top$ y sea $\hat{\mathbf{p}}$ un estimador de $\mathbf{p}$ bajo el diseño complejo. Supongamos ahora que se cumple que

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{\mathcal{D}} N(0, V_{dis}),$$

donde V_{dis} es la varianza asintótica del vector $(X_{11}, X_{12}, \ldots, X_{ab-1})$. Note que si nuestro esquema de muestreo fuese un MAS, $(X_{11}, X_{12}, \ldots, X_{ab-1})$ tendría distribución multinomial y la matriz de varianza-covarianza del diseño tomaría la forma $V_{dis} = P_0 = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$.

Dado que $\boldsymbol{\theta}$ es una función de $\mathbf{p}$, podríamos utilizar el método delta para justificar que

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(0, HV_{dis}H^\top),$$

donde $H = H(\boldsymbol{\theta})$ es la matriz de orden $(a-1)(b-1) \times (ab-1)$ conformada por las derivadas parciales de las componentes de $\boldsymbol{\theta}$ con respecto a las componentes de $\mathbf{p}$.

Uno de lo primeros procedimientos para contrastar independencia en datos categóricos con muestras complejas fue desarrollado por Koch y Freeman (1975). Este test de tipo Wald contrasta la hipótesis de independencia

$$H_0 : \boldsymbol{\theta} = 0$$

usando el estadístico

$$\chi^2_{W0} = \hat{\boldsymbol{\theta}}^\top (\hat{H}\hat{V}_{dis}\hat{H}^\top)^{-1}\hat{\boldsymbol{\theta}},$$

donde $\hat{H} = H(\hat{\boldsymbol{\theta}})$ y $\hat{V}_{dis}$ es un estimador consistente de V_{dis}. Este estadístico tiene asintóticamente una distribución chi-cuadrado con $(a-1)(b-1)$ grados de libertad. Aquí cabe aclarar que si se dispusiera de un estimador consistente de la varianza de $\hat{\boldsymbol{\theta}}$, por alguna técnica de remuestreo, este podría usarse también en lugar de $\hat{H}\hat{V}_{dis}\hat{H}^\top$ para definir el estadístico de tipo Wald.

Un problema con el procedimiento anterior es que si la tabla es grande, el número de unidades primarias debería ser realmente grande como para poder estimar todas las componentes en V_{dis}. Algunos ajustes y procedimientos posteriores que buscan resolver este y otros problemas asociados a este test se revisan en Thomas y Rao (1990).

Una alternativa más usada y eficiente fue formulada a través de los trabajos de Rao y Scott (1984), quienes propusieron corregir el estadístico chi-cuadrado de Pearson multiplicándolo por una constante adecuada. La metodología se basa en el siguiente resultado asintótico

de Rao y Scott (1981). Ellos mostraron que, bajo H_0, el estadístico χ_0^2 de Pearson puede descomponerse como

$$\chi_0^2 = \sum_{i=1}^{(a-1)(b-1)} \lambda_i W_i,$$

donde los $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{(a-1)(b-1)} > 0$ denotan los autovalores de la matriz de diseño generalizada $D = (HP_0H^\top)^{-1}(HV_{dis}H^\top)$, y $W_1, W_2, \ldots, W_{(a-1)(b-1)}$ son variables aleatorias independientes con distribuciones chi-cuadrado de un grado de libertad. La corrección de primer orden sugerida por Rao y Scott postula, entonces, como estadístico de prueba a,

$$\chi_I^2 = \frac{\chi_0^2}{\bar{\delta}},$$

donde:

$$\bar{\delta} = \frac{1}{(a-1)(b-1)} \sum_{i=1}^{(a-1)(b-1)} \delta_i = \frac{Tr(D)}{(a-1)(b-1)}.$$

De esta manera, el estadístico χ_I^2 tiene media $(a-1)(b-1)$ y, aproximadamente, una distribución chi-cuadrado con estos grados de libertad, de no existir mucha variación en los δ_i.

Rao y Scott mostraron posteriormente que de no disponerse de estimaciones de la matriz V_{dis} (y, en consecuencia, de los autovalores en D para poder estimar $\bar{\delta}$), uno podría utilizar estimaciones de los efectos de diseño $\hat{d}_{ij}, \hat{d}_{i.}$ y $\hat{d}_{.j}$ en las estimaciones de $p_{ij}, p_{i.}$ y $p_{.j}$, respectivamente, a fin de obtener la siguiente aproximación:

$$\hat{\bar{\delta}} = \sum_{i=1}^{a}\sum_{j=1}^{b}(1-\hat{p}_{ij})\hat{d}_{ij} - \sum_{i=1}^{a}(1-\hat{p}_{i.})\hat{d}_{i.} - \sum_{j=1}^{b}(1-\hat{p}_{.j})\hat{d}_{.j}.$$

Años después, Thomas y Roberts (1996) derivaron correcciones de segundo orden al estadístico de Pearson con el fin de incorporar la variabilidad de los autovalores en la matriz D. Ellos propusieron el estadístico

$$\chi_{II}^2 = \frac{\chi_I^2}{1+\hat{a}^2},$$

donde $\hat{a}$ representa el coeficiente de variación de los autovalores en la matriz D estimada. Concretamente, utilizando una aproximación de Satterwaite, ellos mostraron que

$$\hat{a}^2 = \sum_{i=1}^{(a-1)(b-1)} \frac{\hat{\delta}_i^2}{(a-1)(b-1)\hat{\bar{\delta}}^2} - 1.$$

Bajo la corrección de segundo orden, el estadístico χ_{II}^2 tiene una distribución asintótica chi-cuadrado con $\frac{(a-1)(b-1)}{1+\hat{a}^2}$ grados de libertad.

Retornando a la parte práctica, es interesante comentar que la librería survey de R posee el comando `svychisq` que realiza las pruebas chi-cuadrado aquí expuestas. El método por

defecto para este análisis es el de Thomas y Roberts (1996) con la corrección de segundo orden. Como ilustración, reconsideremos nuestro problema sobre la relación entre la opinión sobre el alcalde y la tenencia de auto particular en el contexto del muestreo por conglomerados cuando la correlación intraclase es de 1. Los códigos siguientes, como se apreciarán en los resultados, nos proveen de un procedimiento válido para realizar este contraste.

```
cluster = vector()
for (i in 1:500) cluster = c(cluster,i,i)
nuevos.datos = data.frame(Auto, Opinion, cluster)
cluster_design = svydesign(ids=cluster,fpc=rep(5000,1000),data=nuevos.datos)
svychisq(~Auto+Opinion,cluster_design)

##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(~Auto + Opinion, cluster_design)
## F = 3, ndf = 1, ddf = 500, p-value = 0.1
```

5.5.2. Análisis de regresión

En el análisis de regresión lineal múltiple uno busca expresar una v.a. dependiente Y como una función lineal de p variables independientes o predictoras $x_1, x_2, \ldots, x_p$, las cuales se asumirán, como es usual, fijas. El modelo se plantea como

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_p + \epsilon,$$

donde es común asumir que el error ϵ es una variable aleatoria con distribución normal de media 0 y varianza σ^2, y estos errores se asumen independientes para distintas observaciones. Uno de los objetivos centrales de este modelo es estimar el valor medio de Y dado el vector $\mathbf{x} = (x_1, \ldots, x_p)^\top$ de variables independientes en un elemento no observado de la población. Para ello, uno debe contar con mediciones de la variable aleatoria Y para n elementos seleccionados al azar de la población. Dada esta m.a., el modelo puede escribirse como

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ip} + \epsilon_i, \quad \forall i = 1, 2, \ldots, n$$

o matricialmente como

$$\mathbf{Y}_n = \mathbb{X}_n \boldsymbol{\beta} + \mathbf{E}_n,$$

donde $\mathbf{Y}_n$ es un vector columna de orden $n \times 1$; $\mathbb{X}_n$ es una matriz $n \times (p+1)$ cuya primera columna es de unos; $\boldsymbol{\beta}$ es el vector columna de orden $p+1$ de coeficientes de regresión, y $\mathbf{E}_n$ es un vector $n \times 1$ que contiene a los errores ϵ_i.

El método de mínimos cuadrados nos provee de un estimador de $\boldsymbol{\beta}$ que se obtiene de resolver

$$\text{mín} \sum_{i=1}^{n} \epsilon_i^2 = \text{mín} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2, \tag{5.13}$$

siendo su solución

$$\hat{\boldsymbol{\beta}}_{MCO} = (\mathbb{X}_n^\top \mathbb{X}_n)^{-1} \mathbb{X}_n^\top \mathbf{Y}_n.$$

Luego, la estimación buscada del valor medio de Y para un $\mathbf{x}$ dado, al cual llamamos también el hiperplano de regresión, viene dada por

$$\hat{y}_{\mathbf{x}} = [1, \mathbf{x}^\top] \hat{\boldsymbol{\beta}}_{MCO}.$$

En el contexto de una población finita de tamaño N, $\hat{\boldsymbol{\beta}}_{MCO}$ es formalmente un estimador del vector de parámetros $\boldsymbol{\beta}$ que resuelve (5.13), pero para todos los posibles pares

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$$

del vector de variables independientes y la variable dependiente y en la población; esto es de

$$\boldsymbol{\beta}_N = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{Y}_N,$$

donde $\mathbb{X}$ es un matrix $N \times (p+1)$ que contiene a las variables independientes e $\mathbf{Y}_N$ es el vector columna de orden $N \times 1$ que contiene a todas los valores de la variable dependiente en la población. Si el muestreo es complejo y no simple, podríamos naturalmente considerar de manera alternativa al estimador $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$, cuyas componentes resuelvan la siguiente versión ponderada de (5.13):

$$\text{mín} \sum_{i \in \mathcal{S}} \omega_i (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2, \tag{5.14}$$

donde los ω_i son los pesos asociados a la i-ésima unidad seleccionada en la muestra. A esta se le llama precisamente una inferencia basada en el diseño, la cual difiere de la del modelo en el sentido de que la última realiza la inferencia sobre el proceso que pensamos genera a la población real. En tal caso, aun los coeficientes del modelo ajustado para toda la población estarán sujetos a una incertidumbre estadística y se podría pensar que provienen de una superpoblación, de tal manera que se cumpla que cuando $n, N \to \infty$ y $\frac{n}{N} \to c$, para algún $c \in [0, 1[$, $\boldsymbol{\beta}_N \xrightarrow{P} \boldsymbol{\beta}^*$.

Como se sabe, la solución de (5.14) es estándar en el análisis de regresión y se conoce como un estimador de mínimos cuadrados ponderado. Ella viene dada por

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}_n^\top \mathbb{W}_n \mathbb{X}_n)^{-1} \mathbb{X}_n^\top \mathbb{W}_n \mathbf{Y}_n,$$

donde $\mathbb{W}_n = diag(\omega_i)$ es una matriz diagonal de orden n que contiene solo los pesos asociados a cada una de las unidades seleccionadas; $\mathbb{X}_n$ es una matriz $n \times p+1$ que contiene a las

variables predictoras con una primera columna de unos, e $\mathbf{Y}_n$ es un vector columna de orden n que contiene los valores de la variable dependiente, ambos incluyen solo las unidades seleccionadas.

Si bien los estimadores de mínimos cuadrados ponderados poseen una formulación para su varianza, ella no es aquí válida, pues la matriz de pesos $\mathbb{W}_n$ surge de considerar el diseño y no de asumir heterogeneidad como usualmente se plantea para este tipo de estimadores. Para estimar la varianza de $\hat{\boldsymbol{\beta}}$ utilizaremos, al igual que en Wolter (2007), técnicas de linealización. Note, en primer lugar, que nuestro estimador puede escribirse como

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}_n^\top \mathbb{W}_n \mathbb{X}_n)^{-1} \mathbb{X}_n^\top \mathbb{W}_n (\mathbb{X}_n \boldsymbol{\beta} + \mathbf{E}_n) = \boldsymbol{\beta} + (\mathbb{X}_n^\top \mathbb{W}_n \mathbb{X}_n)^{-1} \mathbb{X}_n^\top \mathbb{W}_n \mathbf{E}_n,$$

donde $\mathbf{E}_n = \mathbf{Y}_n - \mathbb{X}_n \boldsymbol{\beta}$.

Consideremos ahora la función $F(\boldsymbol{\omega}) = (\mathbb{X}^\top \mathbb{W} \mathbb{X})^{-1}$, que tiene como argumento al vector $\boldsymbol{\omega}$ de orden $N \times 1$ que define a la matriz de orden $N \times N$, $\mathbb{W} = diag(\boldsymbol{\omega})$. Sea $\boldsymbol{\omega}_n$ un vector $N \times 1$ cuyas componentes contienen los pesos de muestreo de las unidades seleccionadas y valen 0 en caso contrario. Un desarrollo de Taylor de primer orden para esta función F en el vector $\boldsymbol{\omega}_n$ alrededor del vector columna de unos de orden N, $\boldsymbol{\omega}_0 = \mathbf{1}_N$ nos brinda la aproximación

$$F(\boldsymbol{\omega}_n) = (\mathbb{X}_n^\top \mathbb{W}_n \mathbb{X}_n)^{-1} = (\mathbb{X}^\top \mathbb{X})^{-1} + dF_{\boldsymbol{\omega}_0}(\boldsymbol{\omega}_n - \boldsymbol{\omega}_0).$$

Más aún, dado que por propiedad de diferenciación de matrices $dF_{\boldsymbol{\omega}_0}(\boldsymbol{\omega}_n - \boldsymbol{\omega}_0) = -F(\boldsymbol{\omega}_0) dF^{-1}_{\boldsymbol{\omega}_0}(\boldsymbol{\omega}_n - \boldsymbol{\omega}_0) F(\boldsymbol{\omega}_0)$, se cumplirá aproximadamente que

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left((\mathbb{X}^\top \mathbb{X})^{-1} - (\mathbb{X}^\top \mathbb{X})^{-1}((\mathbb{X}_n^\top \mathbb{W}_n \mathbb{X}_n) - (\mathbb{X}^\top \mathbb{X}))(\mathbb{X}^\top \mathbb{X})^{-1}\right) \mathbb{X}_n^\top \mathbb{W}_n \mathbf{E}_n$$

$$= \boldsymbol{\beta} + (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}_n^\top \mathbb{W}_n \mathbf{E}_n - \left(\mathbb{X}^\top \mathbb{X})^{-1}((\mathbb{X}_n^\top \mathbb{W}_n \mathbb{X}_n) - (\mathbb{X}^\top \mathbb{X}))(\mathbb{X}^\top \mathbb{X})^{-1}\right) \mathbb{X}_n^\top \mathbb{W}_n \mathbf{E}_n.$$

Despreciando el último término de esta expresión, se tendrá entonces que aproximadamente

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}_n^\top \mathbb{W}_n \mathbf{E}_n.$$

Así, considerándose que $E(\mathbb{X}_n^\top \mathbb{W}_n \mathbf{E}_n) = \mathbb{X}^\top \mathbf{E} = \mathbf{0}$ (con $\mathbf{E} = \mathbf{Y}_N - \mathbb{X}\boldsymbol{\beta}$), la varianza de este término resulta ser

$$V(\hat{\boldsymbol{\beta}}) = E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top) = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{G}_n (\mathbb{X}^\top \mathbb{X})^{-1},$$

siendo $\mathbb{G}_n = V(\mathbb{X}_n^\top \mathbb{W}_n \mathbf{E}_n)$ una matriz $(p+1) \times (p+1)$ de varianzas-covarianzas cuya entrada rs viene dada por

$$g_{rs} = Cov(\sum_k x_{rk} \epsilon_k \omega_k \delta_k, \sum_\ell x_{s\ell} \epsilon_\ell \omega_\ell \delta_\ell) = \sum_k \sum_\ell x_{rk} x_{s\ell} \epsilon_k \epsilon_\ell Cov(\delta_k, \delta_\ell), \tag{5.15}$$

donde las sumas recorren las distintas etapas o estratos del diseño.

Dado que tanto los residuales $\mathbf{E}_n$ como el término $(\mathbb{X}^\top\mathbb{X})^{-1}$ son usualmente desconocidos, podríamos estimarlos por $\hat{\mathbf{E}}_n = \mathbf{Y}_n - \mathbb{X}_n\hat{\boldsymbol{\beta}}$ y $(\mathbb{X}_n^\top\mathbb{W}_n\mathbb{X}_n)^{-1}$, respectivamente. Ello nos brinda, finalmente, un estimador tipo "sandwich" de la forma

$$\hat{V}(\hat{\boldsymbol{\beta}}) = (\mathbb{X}_n^\top\mathbb{W}_n\mathbb{X}_n)^{-1}\hat{\mathbb{G}}_n(\mathbb{X}_n^\top\mathbb{W}_n\mathbb{X}_n)^{-1}, \tag{5.16}$$

requiriéndose para su término central una estimación $\hat{\mathbb{G}}_n$ de la varianza del vector $\mathbb{X}_n^\top\mathbb{W}_n\hat{\mathbf{E}}_n$, el cual será particular del diseño empleado.

Ejemplo 5.8. *Con el fin de precisar mejor la estimación de la varianza del vector de coeficientes de regresión, pensemos en un diseño estratificado por conglomerados bietápico en el que la i-ésima UPM al interior del estrato $h = 1, 2, \ldots, H$ es seleccionada con una probabilidad π_{hi}, $i = 1, 2, \ldots, N_h$ y la j-ésima USM dentro de la i-ésima UPM del estrato h es seleccionada con probabilidad (condicional) $\pi_{j|hi}$, $j = 1, 2, \ldots, M_{hi}$, siendo M_{hi} el número de USM dentro la i-ésima UPM. Como es usual, asumiremos que el muestreo en cualquier UPM es independiente del muestreo en cualquier otra UPM. En este contexto, la entrada g_{rs} de la matriz $\mathbb{G}_n$ en (5.15) viene dada por*

$$g_{rs} = Cov(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}}\frac{x_{rhij}\epsilon_{hij}\delta_{hi}\delta_{j|hi}}{\pi_{hi}\pi_{j|hi}}, \sum_{h'=1}^{H}\sum_{i'=1}^{N_h}\sum_{j'=1}^{M_{h'i'}}\frac{x_{sh'i'j'}\epsilon_{h'i'j'}\delta_{h'i'}\delta_{j'|h'i'}}{\pi_{h'i'}\pi_{j'|h'i'}})$$

$$= Cov(\sum_{h=1}^{H}\sum_{i=1}^{N_h}Z_{hi|r}\frac{\delta_{hi}}{\pi_{hi}}, \sum_{h'=1}^{H}\sum_{i'=1}^{N_{h'}}Z_{h'i'|s}\frac{\delta_{h'i'}}{\pi_{h'i'}}),$$

donde $Z_{hi|r} = \sum_{j=1}^{M_{hi}}\frac{x_{rhij}\epsilon_{hij}}{\pi_{j|hi}}\delta_{j|hi}$.

Para explicar mejor la expresión anterior, podríamos utilizar la proposición 1.1, al condicionar sobre el arreglo $\boldsymbol{\delta}_1 = [\delta_{hi}]$ de las variables aleatorias indicadoras de inclusión en la primera etapa dentro de los estratos. Ello resulta en

$$g_{rs} = Cov(E(\sum_{h=1}^{H}\sum_{i=1}^{N_h}Z_{hi|r}\frac{\delta_{hi}}{\pi_{hi}} \mid \boldsymbol{\delta}_1), E(\sum_{h'=1}^{H}\sum_{i'=1}^{N_{h'}}Z_{h'i'|s}\frac{\delta_{h'i'}}{\pi_{h'i'}} \mid \boldsymbol{\delta}_1))$$

$$+E(Cov(\sum_{h=1}^{H}\sum_{i=1}^{N_h}Z_{hi|r}\frac{\delta_{hi}}{\pi_{hi}}, \sum_{h'=1}^{H}\sum_{i'=1}^{N_{h'}}Z_{h'i'|s}\frac{\delta_{h'i'}}{\pi_{h'i'}} \mid \boldsymbol{\delta}_1)).$$

Puesto que, por un lado,

$$E(\sum_{h=1}^{H}\sum_{i=1}^{N_h}Z_{hi|r}\frac{\delta_{hi}}{\pi_{hi}} \mid \boldsymbol{\delta}_1) = \sum_{h=1}^{H}\sum_{i=1}^{N_h}E(Z_{hi|r})\frac{\delta_{hi}}{\pi_{hi}} = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\frac{\delta_{hi}}{\pi_{hi}}z_{rhi.},$$

donde $z_{rhi.} = \sum_{j=1}^{M_{hi}}x_{rhij}\epsilon_{hij}$ y ,por otro lado,

$$Cov(\sum_{h=1}^{H}\sum_{i=1}^{N_h}Z_{hi|r}\frac{\delta_{hi}}{\pi_{hi}}, \sum_{h'=1}^{H}\sum_{i'=1}^{N_{h'}}Z_{h'i'|s}\frac{\delta_{h'i'}}{\pi_{h'i'}} \mid \boldsymbol{\delta}_1) = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\frac{\delta_{hi}^2}{\pi_{hi}^2}Cov(Z_{hi|r}, Z_{hi|s}),$$

donde:

$$Cov(Z_{hi|r}, Z_{hi|s}) = \sum_{j=1}^{M_{hi}} \sum_{j'=1}^{M_{hi}} \frac{x_{rhij}\epsilon_{hij}x_{shij'}\epsilon_{hij'}}{\pi_{j|hi}\pi_{j'|hi}} Cov(\delta_{j|hi}, \delta_{j'|hi}),$$

se tiene que

$$g_{rs} = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{i'=1}^{N_h} \frac{z_{rhi.}z_{shi'.}}{\pi_{hi}\pi_{hi'}} Cov(\delta_{hi}, \delta_{hi'}) + \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{1}{\pi_{hi}} \sum_{j=1}^{M_{hi}}\sum_{j'=1}^{M_{hi}} \frac{x_{rhij}\epsilon_{hij}x_{shij'}\epsilon_{hij'}}{\pi_{j|hi}\pi_{j'|hi}} Cov(\delta_{j|hi}, \delta_{j'|hi})$$

$$= \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{z_{rhi.}z_{shi.}(1-\pi_{hi})}{\pi_{hi}} + \sum_{h=1}^{H}\sum_{\substack{i=1\\ i\neq i'}}^{N_h}\sum_{i'=1}^{N_h} \frac{z_{rhi.}z_{shi'.}}{\pi_{hi}\pi_{hi'}}(\pi_{hi,hi'} - \pi_{hi}\pi_{hi'})$$

$$+ \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{1}{\pi_{hi}} \sum_{j=1}^{M_{hi}} x_{rhij}\epsilon_{hij}^2 x_{shij} \frac{(1-\pi_{j|hi})}{\pi_{j|hi}} + \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{1}{\pi_{hi}} \sum_{\substack{j=1\\ j\neq j'}}^{M_{hi}}\sum_{j'=1}^{M_{hi}} \frac{x_{rhij}\epsilon_{hij}x_{shij'}\epsilon_{hij'}}{\pi_{j|hi}\pi_{j'|hi}}(\pi_{j,j'|hi} - \pi_{j|hi}\pi_{j'|hi}).$$

Al igual que en el caso de los estimadores de Horvitz-Thompson, si $\boldsymbol{\beta}$ fuese conocido, un estimador insesgado de g_{rs} vendría dado por

$$\hat{g}_{rs} = \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{z_{rhi.}z_{shi.}(1-\pi_{hi})}{\pi_{hi}^2}\delta_{hi} + \sum_{h=1}^{H}\sum_{\substack{i=1\\ i\neq i'}}^{N_h}\sum_{i'=1}^{N_h} \frac{z_{rhi.}z_{shi'.}}{\pi_{hi}\pi_{hi'}\pi_{hi,hi'}}(\pi_{hi,hi'} - \pi_{hi}\pi_{hi'})\delta_{hi}\delta_{hi'}$$

$$+ \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{1}{\pi_{hi}^2} \left(\sum_{j=1}^{M_{hi}} x_{rhij}\epsilon_{hij}^2 x_{shij} \frac{(1-\pi_{j|hi})}{\pi_{j|hi}^2}\delta_{j|hi} \right) \delta_{hi}$$

$$+ \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{1}{\pi_{hi}^2} \left(\sum_{\substack{j=1\\ j\neq j'}}^{M_{hi}}\sum_{j'=1}^{M_{hi}} \frac{x_{rhij}\epsilon_{hij}x_{shij'}\epsilon_{hij'}}{\pi_{j|hi}\pi_{j'|hi}\pi_{j,j'|hi}}(\pi_{j,j'|hi} - \pi_{j|hi}\pi_{j'|hi})\delta_{j|hi}\delta_{j'|hi} \right) \delta_{hi}.$$

Puesto que $\boldsymbol{\beta}$ es desconocido, este podría estimarse por $\hat{\boldsymbol{\beta}}$, los residuales por $\hat{\mathbf{E}}_n = \mathbf{Y}_n - \mathbb{X}_n\hat{\boldsymbol{\beta}}$ y los términos $z_{rhi.} = \sum_{j=1}^{M_{hi}} x_{rhij}\epsilon_{hij}$, que involucran a los residuales, por $\hat{z}_{rhi.} = \sum_{j=1}^{M_{hi}} x_{rhij}\hat{\epsilon}_{hij}\delta_{j|hi}$. Con ello obtendremos finalmente el estimador

$$\hat{g}_{rs} = \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{\hat{z}_{rhi.}\hat{z}_{shi.}(1-\pi_{hi})}{\pi_{hi}^2}\delta_{hi} + \sum_{h=1}^{H}\sum_{\substack{i=1\\ i\neq i'}}^{N_h}\sum_{i'=1}^{N_h} \frac{\hat{z}_{rhi.}\hat{z}_{shi'.}}{\pi_{hi}\pi_{hi'}\pi_{hi,hi'}}(\pi_{hi,hi'} - \pi_{hi}\pi_{hi'})\delta_{hi}\delta_{hi'} \quad (5.17)$$

$$+ \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{1}{\pi_{hi}^2} \left(\sum_{j=1}^{M_{hi}} x_{rhij}\hat{\epsilon}_{hij}^2 x_{shij} \frac{(1-\pi_{j|hi})}{\pi_{j|hi}^2}\delta_{j|hi} \right) \delta_{hi}$$

$$+\sum_{h=1}^{H}\sum_{i=1}^{N_h}\frac{1}{\pi_{hi}^2}\left(\sum_{\substack{j=1\\ j\neq j'}}^{M_{hi}}\sum_{j'=1}^{M_{hi}}\frac{x_{rhij}\hat{\epsilon}_{hij}x_{shij'}\hat{\epsilon}_{hij'}}{\pi_{j|hi}\pi_{j'|hi}\pi_{j,j'|hi}}(\pi_{j,j'|hi}-\pi_{j|hi}\pi_{j'|hi})\delta_{j|hi}\delta_{j'|hi}\right)\delta_{hi}.$$

Al igual que en los estimadores de Horvitz-Thompson, no resulta difícil probar que el término g_{rs} se puede escribir también como

$$g_{rs}=\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{i'>i}^{N_h}(\pi_{hi}\pi_{hi'}-\pi_{hi,hi'})(\frac{z_{rhi.}}{\pi_{hi}}-\frac{z_{rhi'.}}{\pi_{hi'}})(\frac{z_{shi.}}{\pi_{hi}}-\frac{z_{shi'.}}{\pi_{hi'}})+\sum_{h=1}^{H}\sum_{i=1}^{N_h}\frac{1}{\pi_{hi}}Cov(Z_{hi|r},Z_{hi|s}).$$

Por tanto, un estimador tipo Sen-Yates-Gundy puede implementarse en este caso y viene dado por

$$\hat{g}_{rs}=\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{i'>i}^{N_h}(\frac{\pi_{hi}\pi_{hi'}-\pi_{hi,hi'}}{\pi_{hi,hi'}})(\frac{\hat{z}_{rhi.}}{\pi_{hi}}-\frac{\hat{z}_{rhi'.}}{\pi_{hi'}})(\frac{\hat{z}_{shi.}}{\pi_{hi}}-\frac{\hat{z}_{shi'.}}{\pi_{hi'}})\delta_{hi}\delta_{hi'}+\sum_{h=1}^{H}\sum_{i=1}^{N_h}\frac{1}{\pi_{hi}^2}\hat{Cov}(Z_{hi|r},Z_{hi|s})\delta_{hi},\tag{5.18}$$

donde $\hat{Cov}(Z_{hi|r},Z_{hi|s})$ es una estimación que se realiza sobre la base de las USM seleccionadas al interior de las UPM escogidas. Este es el estimador implementado por defecto en el comando `svyglm`. □

Ejemplo 5.9. *Consideremos la base de datos api y supongamos que deseamos estimar, bajo un muestreo estratificado de conglomerados de una etapa, el rendimiento medio por colegio en el 2000 basándonos en su porcentaje de profesores completamente calificados (full), padres con estudios de posgraduación (grad.sch), estudiantes que están aprendiendo inglés (ell), estudiantes que tienen comidas subsidiadas (meals) y estudiantes para los que este es su primer año en la escuela (mobility). Dado que api es una base de datos poblacional, podríamos teóricamente calcular el vector de parámetros $\boldsymbol{\beta}$ del modelo. Este, descartando los casos perdidos en las variables de interés, viene dado por*

```
data(api)
N0 = dim(apipop)[1]
Pob = apipop[order(apipop$stype,apipop$dnum),]  # apipop ordenado
Pob$cod0 = 1:N0
Pob$b0 = rep(1,N0)
M = as.matrix(na.omit(Pob[,c(38,39,34,32,21,20,23,12)]))
index = as.vector(M[,1]) # indice de casos validos
X = M[,2:7]
Y = M[,8]
beta = solve(crossprod(X))%*%crossprod(X,Y)
beta
```

```
##              [,1]
## b0       600.989
## full       1.753
## grad.sch   2.547
## ell       -0.896
## meals     -1.957
## mobility  -0.101
```

Para la muestra consideraremos como antes el tipo de escuela (stype) como variable de estratificación y los distritos escolares (dnum) como conglomerados. Optaremos por una muestra de, aproximadamente, 30 distritos escolares, los cuales los distribuiremos proporcionalmente a la cantidad de colegios por estrato. Ello nos llevará a consignar 14 colegios elementales, 9 high schools y 7 colegios medios. El diseño y la toma de la muestra se presentan a continuación

```
set.seed(12345)
Pob = Pob[index,]
N1 = dim(Pob)[1]
tt = table(Pob$stype)
ls1 = list(as.vector(tt),c(14,9,7))
Pob$Nh = rep(ls1[[1]],tt)
Pob = cbind(cod = 1:N1,Pob)
mues=mstage(Pob,stage=list("stratified","cluster"),
                  varnames=list("stype","dnum"),
                  size=ls1,method=list("","srswor"),description=T)

## STAGE 1
## Number of strata: 3
## STAGE  2
## Number of selected clusters: 14
## Number of units in the population and number of selected units: 4417 112
## Number of selected clusters: 9
## Number of units in the population and number of selected units: 753 23
## Number of selected clusters: 7
## Number of units in the population and number of selected units: 1018 11

mues = getdata(Pob,mues)[[2]]
dmuesr<-svydesign(id=~dnum, strata=~stype, fpc=~Nh,nest=T,data=mues)
dmuesr
```

```
## Stratified 1 - level Cluster Sampling design
## With (30) clusters.
## svydesign(id = ~dnum, strata = ~stype, fpc = ~Nh, nest = T, data = mues)
```

El análisis de regresión bajo el diseño se realizará con el comando svyglm mediante

```
summary(svyglm(api00~full+grad.sch+ell+meals+mobility, design=dmuesr))

##
## Call:
## svyglm(formula = api00 ~ full + grad.sch + ell + meals + mobility,
##     design = dmuesr)
##
## Survey design:
## svydesign(id = ~dnum, strata = ~stype, fpc = ~Nh, nest = T, data = mues)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  667.021     88.812    7.51  1.7e-07 ***
## full           1.295      0.707    1.83    0.080 .
## grad.sch       2.234      1.031    2.17    0.041 *
## ell           -0.825      0.728   -1.13    0.269
## meals         -2.123      0.728   -2.92    0.008 **
## mobility      -0.411      0.799   -0.51    0.612
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2647)
##
## Number of Fisher Scoring iterations: 2
```

Como se aprecia, solo las proporciones de profesores completamente capacitados y de padres con estudios de posgrado parecen tener significativamente un efecto positivo en el rendimiento medio de la escuela; por otro lado, solo la proporción de estudiantes con comidas subsidiadas pareciera tener un efecto negativo en el rendimiento medio de la escuela. Note que los coeficientes de regresión podríamos haberlos también obtenido en R con

```
Xn = cbind(rep(1,dim(mues)[1]),mues$full,mues$grad.sch,mues$ell,mues$meals,
mues$mobility)
```

```
Yn = mues$api00
w = weights(dmuesr)
Wn = diag(w)
Hn = solve((t(Xn)%*%Wn%*%Xn))
(betah = Hn%*%t(Xn)%*%Wn%*%Yn)

##          [,1]
## [1,] 667.021
## [2,]   1.295
## [3,]   2.234
## [4,]  -0.825
## [5,]  -2.123
## [6,]  -0.411
```

La matriz de varianzas-covarianzas de estos estimadores se pueden también obtener usando las ecuaciones (5.17) y (5.18), que en este diseño se traducen en

$$\hat{g}_{rs,HT} = \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{\hat{z}_{rhi.}\hat{z}_{shi.}(1-\pi_{hi})}{\pi_{hi}^2}\delta_{hi} + \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{\substack{i'=1 \\ i\neq i'}}^{N_h} \frac{\hat{z}_{rhi.}\hat{z}_{shi'.}}{\pi_{hi}\pi_{hi'}\pi_{hi,hi'}}(\pi_{hi,hi'}-\pi_{hi}\pi_{hi'})\delta_{hi}\delta_{hi'}$$

y

$$\hat{g}_{rs,SGY} = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{i'>i}^{N_h}(\frac{\pi_{hi}\pi_{hi'}-\pi_{hi,hi'}}{\pi_{hi,hi'}})(\frac{\hat{z}_{rhi.}}{\pi_{hi}}-\frac{\hat{z}_{rhi'.}}{\pi_{hi'}})(\frac{\hat{z}_{shi.}}{\pi_{hi}}-\frac{\hat{z}_{shi'.}}{\pi_{hi'}})\delta_{hi}\delta_{hi'},$$

donde $\pi_{hi} = \frac{n_h}{N_h}$ *y* $\pi_{hi,hi'} = \frac{n_h(n_h-1)}{N_h(N_h-1)}$. *Una evaluación de los errores estándares de estimación estimados, a partir de estos estimadores, se muestra en la siguiente tabla:*

Parámetro	β_0 *Intercepto*	β_1 *full*	β_2 *grad.sch*	β_3 *ell*	β_4 *meals*	β_5 *mobility*
Estimado	*667.0213*	*1.2951*	*2.2344*	*-0.8253*	*-2.1226*	*-0.4114*
Std.Error (HT)	*89.1709*	*0.7025*	*1.0434*	*0.7199*	*0.7391*	*0.7861*
Std.Error (SGY)	*88.8118*	*0.7068*	*1.0307*	*0.7282*	*0.7276*	*0.7993*

Como se observa, las estimaciones de los errores estándar para los coeficientes son muy similares y la última coincide con el de la salida del comando `svyglm`.

Si no consideráramos los pesos de muestreo, el análisis nos brindaría la siguiente salida:

```
summary(glm(api00~full+grad.sch+ell+meals+mobility, data=mues))
```

```
##
## Call:
## glm(formula = api00 ~ full + grad.sch + ell + meals + mobility,
##     data = mues)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -217.34   -35.74     0.76    38.25   165.16
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  562.584     50.800   11.07   <2e-16 ***
## full           1.858      0.501    3.71   0.0003 ***
## grad.sch       3.723      0.692    5.38    3e-07 ***
## ell           -1.459      0.526   -2.77   0.0063 **
## meals         -0.991      0.394   -2.51   0.0130 *
## mobility      -0.556      0.485   -1.15   0.2536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3722)
##
##     Null deviance: 1718868  on 145  degrees of freedom
## Residual deviance:  521065  on 140  degrees of freedom
## AIC: 1623
##
## Number of Fisher Scoring iterations: 2
```

Se aprecia, entonces, una mínima diferencia en ambos análisis. Note también los menores errores de estimación de los estimadores de mínimos cuadrados. □

Se sabe que la extensión del análisis de regresión lineal múltiple para otro tipo de respuestas, como binarias, de conteo o no negativas, puede realizarse a través de los modelos lineales generalizados. En estos, el método de estimación no es el de mínimos cuadrados sino de máxima verosimilitud. Este método requiere la maximización de la función de probabilidad o densidad conjunta de las respuestas, o de su logaritmo, las últimas que se asumen que son independientes y que se asocian al predictor lineal mediante funciones pre definidas de enlace g que dependen del tipo de respuesta. Se asume que el modelo lineal general de trabajo en cuestión pertenece a una familia exponencial que relaciona para una observación

i su media o media condicional μ con un predictor lineal mediante

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

y que su varianza marginal puede ser razonablemente aproximada por

$$V(Y \mid \mathbf{X} = \mathbf{x}_i) = \sigma^2 V(\mu_i).$$

En el caso, por ejemplo, de la regresión logística para respuestas binarias, la función de log-verosimilitud viene dada por

$$l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\pi(\mathbf{x}_i)) + (1 - y_i) \log(1 - \pi(\mathbf{x}_i)),$$

donde $\pi(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} = \mu_i$ representa el valor que se espera tome la variable respuesta binaria Y y corresponde a la inversa de la función de enlace logístico $g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i}) = \mathbf{x}_i^\top \boldsymbol{\beta}$.

En el muestreo complejo, la asunción de independencia entre las distintas respuestas raramente se cumple y, por tanto, este procedimiento podría resultar inválido. Para suplir ello se ha propuesto en la literatura una metodología de pseudo máxima verosimilitud asistida por el modelo que incorpora los pesos de muestreo a la función última. La función de log-pseudo-máxima verosimilitud a optimizar en la regresión logística es

$$l_P(\boldsymbol{\beta}) = \log PL(\boldsymbol{\beta}) = \sum_{i \in \mathcal{M}} \omega_i (y_i \log(\pi(\mathbf{x}_i)) + (1 - y_i) \log(1 - \pi(\mathbf{x}_i))).$$

Una vez obtenidos los estimadores, la estimación de sus varianzas pueden obtenerse ya sea por los métodos de linealización o replicación estudiados.

En R, el procedimiento para el ajuste e inferencia de estos modelos se encuentra implementado en el comando `svyglm` del paquete survey.

En este texto introductorio no discutiremos al detalle todos los modelos de regresión lineal generalizados ni su análisis de ajuste, desarrollos que cabe comentar son en muchos casos aún temas de actual investigación. Para mayores detalle, el lector interesado puede consultar el texto de Heeringa y Berglund (2010) y el artículo de Binder (1983). Lo que sí vale la pena comentar es el proceso de inferencia. En general, si estamos interesados en un IC al 100(1 - α) % o en una prueba de significación sobre cualesquieras de los coeficientes de regresión β_i del modelo, estos vienen caracterizados por

$$[\hat{\beta}_i - t_{1-\frac{\alpha}{2}}(gl)\hat{SE}(\hat{\beta}_1), \hat{\beta}_i + t_{1-\frac{\alpha}{2}}(gl)\hat{SE}(\hat{\beta}_1)]$$

y la estadística de prueba $t = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}$, la cual tiene distribución $t-$ de Student con $gl = \sum_h a_h - H$ grados de libertad, siendo el primer término el número de conglomerados y el segundo el número de estratos, ambos de la primera etapa del diseño. Naturalmente, $t_{1-\frac{\alpha}{2}}(gl)$ denota aquí el cuantil $1 - \frac{\alpha}{2}$ de la distribución t de Student.

Por otro lado, las conocidas pruebas F sobre un grupo de regresores pueden sustituirse por las pruebas de Wald a través del estadístico

$$F_W = \frac{1}{q}\hat{\boldsymbol{\beta}}_q^\top \hat{\Sigma}_q^{-1}\hat{\boldsymbol{\beta}}_q,$$

donde $\hat{\boldsymbol{\beta}}_q$ denota el estimador de cualquier vector de coeficientes de regresión estimados de dimensión $1 \leq q \leq p$ que sean un subconjunto del vector de todos los p coeficientes de regresión en el modelo y $\hat{\Sigma}_q$ es su correspondiente matriz de varianzas-covarianzas estimada. Bajo la hipótesis nula $H_0 : \boldsymbol{\beta}_q = 0$, se cumple que asintóticamente F_W tiene distribución F de Fisher con q grados de libertad en el numerador y gl grados de libertad en el denominador. Todas estas pruebas se encuentran implementadas en R bajo el comando `regTermTest` del paquete survey.

Lumley y Scott (2014) argumentan, sin embargo, que en lugar de las pruebas de Wald sería preferible usar pruebas de razón de verosimilitud, ya que estas son, a diferencia de las primeras, invariantes a transformaciones de los parámetros y muestran mejores propiedades en muestras pequeñas. Para ello, ellos extienden las pruebas de Rao y Scott vistas en el capítulo anterior a un contexto mucho más general. Recordemos que en las pruebas de razón de verosimilitud es de interés particionar el vector de parámetros $\boldsymbol{\beta}$ de dimensión p como $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, donde $\boldsymbol{\beta}_0$ tiene dimensión $q \leq p$ y contrastar la hipótesis nula $H_0 : \boldsymbol{\beta}_0 = 0$. En este modelo más general, la función de pseudo-log-verosimilitud viene dada por

$$l_P(\boldsymbol{\beta}) = \sum_{i\in\mathcal{S}} \omega_i \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}),$$

donde $f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta})$ denota la función de densidad o probabilidad de las respuestas en función de las covariables y del vector de parámetros. Si $\hat{\boldsymbol{\theta}}_0$ denota la la solución de

$$U(\boldsymbol{\beta}) = \frac{\partial l_P(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i\in\mathcal{S}} \omega_i \frac{1}{g'(\mu_i)V(\mu_i)}(y_i - \mu_i) = 0$$

bajo la restricción que $\boldsymbol{\beta}_0 = 0$, Lumley y Scott (2014) muestran que bajo H_0 y ciertas condiciones de regularidad se cumple que, conforme $n, N \to \infty$,

$$\Lambda_n = 2(l_P(\hat{\boldsymbol{\theta}}) - l_P(\hat{\boldsymbol{\theta}}_0)) \xrightarrow{\mathcal{D}} \sum_{i=1}^{q} \delta_i Z_i^2,$$

donde $Z_1, Z_2, \ldots Z_q$ es una m.a. de variables normales estándar independientes y $\delta_1, \delta_2, \ldots, \delta_q$ son los autovalores de la matriz de $\Lambda = (I_{11} - I_{12}I_{22}^{-1}I_{21})V_{11}$ en las que V_{11} denota la matriz de varianza-covarianza asintótica de $\sqrt{n}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*)$ e

$$I(\boldsymbol{\beta}^*) = E(-\frac{\partial^2 l_P(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^\top}) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{11} \end{bmatrix}.$$

Estos autores muestran también una aproximación de Satterthwaite para la distribución de Λ dada por $\frac{\Lambda}{\delta} \sim \chi^2(\nu)$, con $\nu = \frac{\sum_{i=1}^q \delta_i^2}{(\sum_{i=1}^q \delta_i)^2}$, recomendable cuando los δ_i son muy heterogéneos.

Ejemplo 5.10. *Consideremos el siguiente ejemplo tomado del material suplementario que acompaña al texto de Fox y Weisberg (2018), donde es de interés estudiar la actitud de las personas hacia la prohibición del aborto. Para ello consideraremos la CES 2011, la cual fue una encuesta electoral realizada durante el período de campaña 2011 en Canadá. Esta base de datos se encuentra disponible en los paquetes car y carData de R que acompañan al texto de Fox y Weisberg (2018). Mayores detalles sobre la encuesta se pueden encontrar en Fournier y Stolle (2013). En el CES 2011 el país fue dividido en estratos conformados por las 10 provincias de Canadá. Dentro de cada estrato h se tomó un MASs basándose en un marco muestral de los números telefónicos de los hogares. Dado que las provincias en Canadá son bastante distintas en tamaño y con vistas a facilitar la comparación entre provincias, las provincias más pequeñas fueron sobremuestradas. Como consecuencia, diferentes hogares tuvieron una desigual aunque conocida probabilidad de selección dentro de la muestra. Por otro lado, los hogares seleccionados fueron contactados por teléfono y allí se determinó el número de votantes elegibles en cada hogar. En una segunda etapa de muestreo solo un individuo fue seleccionado al azar entre los individuos elegibles en el hogar. Por tal razón, los individuos que pertenecen a hogares numerosos tendrán una menor probabilidad de ser seleccionados en la muestra que aquellos que viven en hogares pequeños. La base de datos CES11 posee la siguiente estructura:*

```
library(carData)
library(car)
data("CES11")
CES11[1:8,2:9]

##   province population weight gender abortion importance education urban
## 1       BC    3267345   4288 Female       No   somewhat    somePS urban
## 2       QC    5996930   9231   Male       No        not bachelors urban
## 3       QC    5996930   6154   Male      Yes   somewhat   college urban
## 4       NL     406455   3430 Female       No       very    somePS urban
## 5       ON    9439960   8978   Male       No        not    higher rural
## 6       ON    9439960   8978 Female       No        not    higher urban
## 7       NL     406455   3430 Female      Yes       very    lessHS rural
## 8       NL     406455   1715 Female      Yes    notvery   college urban
```

En orden correlativo, la primera columna de la base de datos CES11 identifica al hogar seleccionado, la segunda al estrato o provincia (province), la tercera al número de individuos en la provincia donde el entrevistado vive (population), la cuarta al peso de muestreo (weight), la quinta al sexo del individuo entrevistado (gender), la sexta a su respuesta a la pregunta central del estudio: ¿Debería prohibirse el aborto? (abort), la séptima a su calificativo en escala de Likert de la importancia que le da a religión en su vida (importance), la octava a

su nivel de educación (education) y la última a si vive en una zona rural o urbana (urban).

Antes de realizar una regresión binaria sobre la actitud de la población canadiense hacia el aborto en función del género, nivel de educación, zona donde vive (rural o urbana) e importancia dada a la religión, sería interesante describir primero cómo se comporta nuestra variable dependiente. El siguiente código muestra este análisis y la definición del diseño.

```
dCES11 = svydesign(ids=~id,strata = ~province, fpc=~population,
weights = ~weight, data=CES11)
svymean(~abortion,design=dCES11)

##               mean   SE
## abortionNo  0.815 0.01
## abortionYes 0.185 0.01
```

Como se ve, el 81.5 % de las personas encuestadas se oponen a prohibir el aborto.

Comparemos ahora los análisis de regresión logística, bajo el diseño y el modelo, en términos de sus coeficientes y valores p

```
dreg = svyglm(abortion~importance+gender+education+urban, design=dCES11,
family=quasibinomial)
mreg = glm(abortion ~ importance + gender  + education + urban, data=CES11,
family=binomial)
compareCoefs(dreg,mreg,zvals=T,pvals=T)

## Calls:
## 1: svyglm(formula = abortion ~ importance + gender + education + urban,
##    design = dCES11, family = quasibinomial)
## 2: glm(formula = abortion ~ importance + gender + education + urban,
##   family = binomial, data = CES11)
##
##                     Model 1 Model 2
## (Intercept)          -3.578  -3.446
## SE                    0.324   0.280
## z                    -11.03  -12.30
## Pr(>|z|)           < 2e-16 < 2e-16
##
## importancenotvery    0.458   0.442
## SE                   0.348   0.310
## z                     1.32    1.43
## Pr(>|z|)             0.1880  0.1539
```

```
##
## importancesomewhat   1.327   1.203
## SE                   0.271   0.235
## z                     4.89    5.12
## Pr(>|z|)           1.0e-06 3.1e-07
##
## importancevery       3.141   2.977
## SE                   0.262   0.225
## z                    12.00   13.21
## Pr(>|z|)           < 2e-16 < 2e-16
##
## genderMale           0.328   0.375
## SE                   0.148   0.127
## z                     2.21    2.95
## Pr(>|z|)            0.0270  0.0032
##
## educationcollege     0.418   0.393
## SE                   0.229   0.198
## z                     1.83    1.99
## Pr(>|z|)            0.0676  0.0468
##
## educationhigher     0.3048 -0.0359
## SE                  0.2994  0.2642
## z                     1.02   -0.14
## Pr(>|z|)            0.3087  0.8920
##
## educationHS          0.536   0.579
## SE                   0.230   0.194
## z                     2.33    2.99
## Pr(>|z|)            0.0198  0.0028
##
## educationlessHS      0.980   0.901
## SE                   0.250   0.208
## z                     3.92    4.32
## Pr(>|z|)           8.9e-05 1.5e-05
##
## educationsomePS      0.128   0.250
## SE                   0.282   0.234
```

```
## z                      0.45    1.07
## Pr(>|z|)             0.6501  0.2859
##
## urbanurban           -0.283  -0.306
## SE                    0.166   0.136
## z                     -1.70   -2.25
## Pr(>|z|)             0.0885  0.0241
##
```

Como se aprecia, las estimaciones obtenidas son bastante similares. Manteniendo los otros predictores fijos, se aprecia que la oposición al aborto se incrementa con la mayor importancia que se le dé a la religión; esta, además, es mayor en hombres que en mujeres y, en general, mayor en los niveles educativos más bajos, aunque no monótonamente. Finalmente, la oposición a prohibir el aborto es marginalmente más baja en residentes urbanos que en rurales. □

5.5.3. Contrastes de medias para una, dos o más poblaciones.

Los contrastes paramétricos clásicos de medias para una, dos o más poblaciones se realizan con las conocidas estadísticas t, normales y F, las cuales involucran a las medias y varianzas muestrales de las variables de interés en el estudio. Si bien, en un diseño complejo, podríamos adaptar tales estadísticas incorporando la varianza de la media bajo el diseño y ajustando sus grados de libertad, resulta mucho más práctico utilizar más bien un enfoque de regresión y las pruebas de Wald vistas en la subsección 5.5.2. Esta es precisamente la estrategia empleada por el paquete survey a través de su comando `svyttest,` el cual nos permite contrastar la hipótesis nula de que la media de la población toma un valor preespecificado μ_0 o que la media de dos poblaciones es o no la misma.

Ejemplo 5.11. *Suponga que para el diseño del ejemplo 5.9 sea de interés analizar la hipótesis de trabajo que el rendimiento medio del índice api 2000 es significativamente distinto al de 1999. Ello se podría realizar mediante el comando* `svyttest` *o, alternativamente, con el comando* `svyglm` *como seguidamente se muestra*

```
svyttest(I(api00-api99)~0,dmuesr)

##
##  Design-based one-sample t-test
##
## data:  I(api00 - api99) ~ 0
## t = 7, df = 30, p-value = 3e-07
```

```
## alternative hypothesis: true mean is not equal to 0
## sample estimates:
## mean
##   36

summary(svyglm(api00-api99~1, design=dmuesr))

##
## Call:
## svyglm(formula = api00 - api99 ~ 1, design = dmuesr)
##
## Survey design:
## svydesign(id = ~dnum, strata = ~stype, fpc = ~Nh, nest = T, data = mues)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.01       5.29    6.81  2.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 803)
##
## Number of Fisher Scoring iterations: 2
```

Ambos análisis, descartando redondeos, brindan los mismos resultados y muestran que sí existen diferencias significativas entre las medias de los índice api 1999 y 2000.

Otra hipótesis de interés es que el índice api00 este relacionado con el nivel de educación promedio de los padres de los alumnos en estos colegios. Para esto, recordemos que la variable `avg.ed` *recoge el número promedio de años de estudios de los padres en cada colegio. Para simplificar, supongamos que realizamos una clasificación del nivel educativo de los padres por colegio, asignando a cada colegio solo una de 3 categorías creadas al segmentar los puntajes promedios en 3 intervalos de más o menos igual longitud. La distribución de frecuencias y el análisis de esta variable, que llamaremos* `Ed`*, se muestra a continuación:*

```
table(cut(Pob$avg.ed,3))

##
## (0.996,2.33]  (2.33,3.67]     (3.67,5]
##         1771         3478          761

dmuesr = update(dmuesr,Ed = cut(avg.ed,3))
```

Si bien la prueba correspondiente es, formalmente, un ANOVA, sabemos que esta se puede también desarrollar desde un enfoque de regresión, como

```
summary(svyglm(api00~Ed, design=dmuesr))

##
## Call:
## svyglm(formula = api00 ~ Ed, design = dmuesr)
##
## Survey design:
## update(dmuesr, Ed = cut(avg.ed, 3))
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      596.8       38.9   15.34  3.1e-14 ***
## Ed(2.08,3.15]     64.7       41.7    1.55     0.13
## Ed(3.15,4.23]    211.3       42.2    5.01  3.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 5402)
##
## Number of Fisher Scoring iterations: 2
```

Así, solo se aprecian diferencias significativas entre el rendimiento medio de los colegios que tienen padres con un alto nivel educativo en comparación con colegios en los que los padres tienen un bajo nivel. Si bien los padres con niveles altos o intermedios de educación tienen una influencia positiva en el rendimiento de las escuelas, la diferencia de rendimientos entre los colegios con padres de nivel educativo intermedio y bajo es muy marginal y no significativa. Esto también puede apreciase al pedir un reporte del rendimiento de las escuelas según el nivel educativo de los padres.

```
svyby(~api00,~Ed,dmuesr,svymean)

##                          Ed api00   se
## (0.997,2.08] (0.997,2.08]   597 38.9
## (2.08,3.15]   (2.08,3.15]   662 27.3
## (3.15,4.23]   (3.15,4.23]   808 15.0
```

□

5.6. Ejercicios

1. En cierto estudio se empleó un diseño complejo a fin de estimar, entre otras cosas, el índice de pobreza de una región. Para ello se seleccionaron, bajo este diseño, 10 familias, cuyos índices de pobreza y respectivos pesos base de muestreo se muestran a continuación:

Índice	34.8	49.7	23.8	65.4	55.2	38.8	43.7	44.8	59.7	60.3
Peso	167.10	68.04	22.31	167.10	419.81	120	100	54.31	22.54	58.79

Un objetivo del estudio fue determinar los cuartiles de pobreza en esta población. Estime tales cuartiles basándose en los datos tomados y el diseño empleado.

2. Un embarque contiene 60 containers, los cuales transportan un total de 6000 cajas de fruta. Para inspeccionar este embarque se decidió, en una primera etapa, seleccionar al azar y con reemplazamiento 4 containers y luego, de cada container, seleccionar al azar y sin reemplazamiento 3 cajas en las que se registrará el peso de las cajas. Si el muestreo arrojó los siguientes resultados:

Container seleccionado	Número de cajas en el container	Peso (en kg) de las cajas en los containers seleccionados
23	100	10.3, 12.2, 9.8
12	80	11.2, 13.1, 9.9
8	114	8.95, 15.3, 14.4
44	93	11.60, 10.53, 11.8

a) Muestre que la probabilidad de que un container cualquiera sea seleccionado en esta inspección es igual a $1 - (\frac{59}{60})^4$.
b) Halle los pesos de muestreo para cada caja seleccionada en la muestra.
c) Estime el peso promedio de todas las cajas de este embarque.
d) Estime el error estándar de estimación de la estimación en c).
e) Estime el tercer cuartil de los pesos de todas las cajas de este embarque. Muestre en este caso un código que le permita hallar este cuartil sin usar el paquete survey.

3. Considere el diseño de la población penal dado en la sección 4.13.
a) Tome la muestra correspondiente y halle los pesos de muestreo para cada unidad seleccionada.
b) Una de las variables importantes en esta población es la situación jurídica del interno. Estime, bajo este diseño, la proporción de internos sentenciados y el efecto de diseño correspondiente.
c) Ajuste, de ser posible, los pesos en b) por no respuesta y estime, bajo estas nuevas ponderaciones, la proporción de internos sentenciados.

4. En data de dominio público es común, por cuestiones de confidencialidad u otros, no reportar la información completa del diseño y tan solo presentar (pseudo) estratos o conglomerados, cuyo análisis válido solo podrá hacerse a través de los pesos de muestreo consignados. Como ejemplo consideremos la National Health Interview Survey del 2013, encuesta nacional de salud por entrevistas realizada en Estados Unidos. Una versión abreviada de ella se encuentra en el archivo nhis.large del paquete PracTools del libro de Valliant et al. (2013). Esta contiene información de 18 variables sobre un total de 21 588 registros (personas) que respondieron a la encuesta de salud. Más información sobre esta encuesta se encuentra en

`https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm.`

Al no tenerse información precisa sobre este diseño, podríamos considerar que la data proviene de un diseño por conglomerados unietápico estratificado con 2 UPM por estrato. Un aspecto de esta base es que hay varios casos sin respuesta para la variable inc.grp, la cual reporta el ingreso categorizado de la familia del encuestado en dólares. Puesto que el porcentaje de casos perdidos para esta variable podría ser alto, sería de interés ver cómo realizar los ajustes de los pesos estimando las probabilidades de no respuesta.
a) Incluyendo solo a personas menores de 18 años, estime, mediante una regresión logística, las probabilidades de no respuesta para la variable de ingresos. Asuma que los pesos dados son los pesos base y utilice 5 grupos para los ajustes.
b) Estime la distribución etárea en esta población, y para estimar sus errores estándar de estimación utilice el método de linealización y todos los métodos de remuestreo estudiados.

5. Para la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) del 2014 llevada a cabo en México se empleó un diseño probabilístico, estratificado y por conglomerados polietápico. Las unidades primarias de muestreo fueron constituidas por agrupaciones de viviendas con características diferenciadas dependiendo del estrato donde se ubicaban, las unidades secundarias fueron las viviendas y la unidad de observación fue el hogar. Determine el número de hogares a considerar en la ENIGH-2014 si se estimó que el número de hogares para el 2014 en México fue de 31 374 724. Para ello considere como variable de referencia al ingreso corriente total del hogar y que se desea estimar este con un error no mayor a los 1,068 pesos a un nivel de confianza del 90 %. Tome en cuenta que en la ENIGH-2012 se utilizó un diseño similar y que en ella se estimó la desviación estándar de los ingresos por hogar en 44 157.8329 pesos, un efecto de diseño de 3.73 y una tasa de no respuesta del 15 %.

6. Se desea realizar una encuesta por muestreo complejo para averiguar, entre otras cosas, con un error no mayor a 0.07 y una confianza del 95 %, la proporción de satisfacción de los trabajadores en su centro laboral para un determinado sector. Un estudio similar se realizó el año pasado, y para este se reportó un efecto de diseño de 2.4 y un porcentaje de satisfacción del 73 %. Si el marco actualizado de trabajadores del sector cuenta con 12 378 trabajadores, ¿cuál debería ser el tamaño de muestra a considerar?

7. Suponga que en el Casen 2011 estuviéramos interesados en estimar la tasa de pobreza por región con un error de estimación de 0.5 puntos porcentuales menor al tomado en el diseño Casen 2011. Calcule los tamaños de muestra que deberían tomarse para esta nueva especificación. Utilice un nivel de confianza del 95 %.

8. Considere el problema 19 del capítulo 4 en el que una muestra de 4 tiendas de la cadena ha sido seleccionada, con probabilidades proporcionales al stock, y se han obtenido las siguientes ventas del celular YTRON:

Tienda	Núm. de celulares en stock	Núm. de celulares vendidos	Total de ventas del celular	Núm. de celulares vendidos con rebaja
1	55	22	15 395	17
6	120	60	44 230	35
9	35	18	13 440	6
13	30	19	13 470	13

¿Cuáles serían las estimaciones pedidas?

9. En el siguiente ejemplo tomado de Lohr (2000) es de interés estimar la edad promedio de los árboles de un parque. La manera más precisa de medir la edad de un árbol es contando el número de anillos de crecimiento en su tronco. Este método, sin embargo, es complicado, por lo cual alternativamente podríamos simplemente medir el diámetro del árbol. Un guardabosques ha tomado la medida del diámetro de todos los 1132 árboles del parque y encontrado una media de 10.3 pulgadas. Si, por otro lado, seleccionó al azar y sin reemplazamiento 20 árboles del parque para realizar la medición clásica y encontró lo siguiente:

Árbol N^o.	1	2	3	4	5	6	7	8	9	10
Diámetro	12.0	11.4	7.9	9.0	10.5	7.9	7.3	10.2	11.7	11.3
Edad	125	119	83	85	99	117	69	133	154	168
Árbol N^o.	11	12	13	14	15	16	17	18	19	20
Diámetro	5.7	8.0	10.3	12.0	9.2	8.5	7.0	10.7	9.3	8.2
Edad	61	80	114	147	122	106	82	88	97	99

a) Muestre un diagrama de dispersión del diámetro de los árboles seleccionados y la edad obtenida por el método de los anillos.
b) Estime, sin tomar en cuenta el diámetro, la edad media de los árboles del parque.
c) Utilice un estimador de razón para estimar la edad media de los árboles del parque.
d) Ajuste un modelo de regresión entre la edad y el diámetro y use este modelo para la estimación de interés. Compare esta estimación con la de las anteriores.
e) Use el método Jackknife para estimar la varianza del coeficiente de regresión en d). ¿Cómo se compara este estimador con el obtenido con el método de linealización?

10. Un parámetro de interés en la base de datos muestral `apistrat` del paquete survey es el número total de estudiantes que tomaron el test api (variable `api.stu`).

a) Obtenga la estimación de Horvitz-Thompson del número de estudiantes que tomaron el test api, así como su error estándar de estimación.

b) Obtenga la estimación en a) con los métodos jackknife y bootstrap.

c) Obtenga a), pero ahora con un estimador de razón que utilice como variable auxiliar el número total de alumnos matriculados.

d) Indique cuál de las dos estimaciones anteriores elegiría y por qué. Más aún, dado que conoce la base de datos poblacional, compare también las dos estimaciones con el verdadero valor del parámetro de interés.

11. En la subsección 5.4.2 describimos el estimador de razón y una aproximación de la varianza de este estimador vía linealización.

a) Indique cuál cree que sea la la razón por la que el paquete survey, en la estimación de una media bajo un muestreo por conglomerados bietápico, utiliza por defecto para la estimación del error estándar la fórmula de un estimador de razón.

b) Muestre la validez de la fórmula de la varianza de un estimador de razón de la media bajo un diseño por conglomerados bietápico.

12. Murgia (2018) realizó el primer estudio de adopción BIM en proyectos de edificación en Lima y Callao. El BIM es un nuevo método de trabajo que integra a todos los agentes que intervienen en un proceso de edificación, como arquitectos, ingenieros, constructores, promotores, etc., y establece un flujo de comunicación trasversal entre ellos, generando un modelo virtual que incluye toda la información relacionada con la edificación a lo largo de su ciclo de vida. El estudio se basó en un muestreo por conglomerados bietápico, tomándose como marco muestral el 22^o. estudio "El Mercado de las Edificaciones Urbanas de Lima Metropolitana y El Callao"(CAPECO, 2017) y una proyección y actualización del número de obras para el tiempo de recolección de los datos. El marco muestral, unidades primarias seleccionadas y unidades secundarias consideradas se muestran en la figura 5.2. Como se aprecia, el universo de edificaciones en Lima y Callao se dividió en siete sectores urbanos (estratos), que, a su vez, se subdividieron en distritos y zonas. Como se puede ver, Miraflores y San Isidro se ubican en el sector urbano Lima Top. Sin embargo, Miraflores está dividido en 2 zonas al igual que San Isidro. Los conglomerados de primera etapa fueron las zonas y los de segunda las obras en cada una de las zonas. Si las repuestas en la encuesta acerca de si en la obra se había adoptado o no la metodología BIM fueron, respectivamente, para cada una de las obras mostradas en la figura 5.2 las siguientes: No No No Sí No Sí No No No Sí No No No No No No No Sí Sí Sí No No Sí No No Sí No Sí Sí No No No No Sí No Sí No No No Sí Sí No No No No No Sí Sí No No No Sí No No Sí No No No No No No No Sí No No No No No No No No Sí No No Sí No No No No No No No No No Sí No No No Sí No No No No No No No No No Sí No No No Sí No No No No Sí No No No No No No No No No No No No No No Sí No No Sí No No No Sí No No No No No Sí No No No Sí No Sí No Sí No No

Sí No No Sí Sí No Sí No No Sí Sí Sí No No Sí No No No Sí Sí No No No No No No Sí Sí No No
No Sí No No No No No No No Sí No No No No No No No No Sí Sí Sí No No No No No No No
No No No No No No No No No No No No No No No No No No No No No No No No No No No
No No No No No No No No No No No Sí No No No No No No No No No No No No No Sí No No
No No Sí Sí No No No No No No No Sí Sí No Sí No No No No Sí No No No Sí Sí Sí Sí Sí Sí No
Sí Sí No No No No No No No No No No No No No No No No No No

Estime el porcentaje de obras que adoptaron el BIM en Lima y Callao durante el 2017, así como su margen de error al 95 % de confianza.

LOCALIZACIÓN			Obras	Encuestas
ESTRATO	DISTRITO	ZONA	por zona	por zona
Lima Top	Miraflores	A	54	4
	Miraflores	B	77	21
	San Isidro	A	77	20
	San Isidro	B	22	2
	La Molina	A	5	
	La Molina	B	2	
	Santiago de Surco	A	104	21
	Santiago de Surco	B	69	20
	Santiago de Surco	C	20	
	San Borja	A	17	17
	San Borja	B	51	6
	Barranco	A	29	10
	Barranco	B	35	7
Lima Moderna	Jesús María	A	17	3
	Jesús María	B	12	2
	Jesús María	C	3	2
	Lince	A	7	2
	Lince	B	32	6
	Magdalena	A	20	20
	Magdalena	B	11	3
	Magdalena	C	17	4
	Pueblo Libre	A	50	13
	Pueblo Libre	B	17	
	Pueblo Libre	C	27	
	San Miguel	A	8	5
	San Miguel	B	12	12
	San Miguel	C	44	
	Surquillo	A	9	5
	Surquillo	B	13	3
	Surquillo	C	12	2
Callao	Bellavista	A	25	4
	Callao	A	8	8
	La Perla	A	6	
	Ventanilla	A	3	2

LOCALIZACIÓN			Obras	Encuestas
ESTRATO	DISTRITO	ZONA	por zona	por zona
Lima Centro	Cercado de Lima	A	17	10
	Cercado de Lima	B	10	
	Breña	A	15	11
	Breña	B	4	
	Breña	C	6	
	La Victoria	A	20	3
	San Luis	A	3	
Lima Este	Ate	A	50	7
	Cieneguilla	A	1	
	Chaclacayo	A	3	
	Lurigancho	A	6	
	Santa Anita	A	10	3
	El Agustino	A	9	3
	San Juan de Lurigancho	A	19	
Lima Norte	Carabayllo	A	23	3
	Cómas	A	42	6
	Los Olivos	A	10	4
	Los Olivos	B	44	8
	Los Olivos	C	3	
	Puente Piedra	A	4	
	San Martín de Porres	A	6	4
	Independencia	A	1	
	Ancón	A	3	
Lima Sur	Chorrillos	A	3	3
	Chorrillos	B	42	10
	Lurin	A	5	
	Pachacamac/VES	A	24	2
	San Juan de Miraflores	A	23	8
	Pucusana	A	4	
	Punta Hermosa/Negra	A	14	2
	San Bartolo	A	9	8
	Santa María del Mar	A	4	

Figura 5.2: Estratos y conglomerados del estudio BIM

13. Suponga que en el 2016 el gobierno de la región de Cajamarca tenía interés en conocer, entre otras cosas, la proporción de internos sentenciados, de reos que consumían drogas y la distribución de tipos de delito cometidos por los internos de sus penales. Para ello se diseñó una encuesta por muestreo que consideró a cada establecimiento penal como un dominio. El diseño definió como variable de estratificación al género de los internos del penal. Por otro lado, dado que los penales en la región son bien diferenciados, siendo de lejos el de Cajamarca el más grande (los otros dominios son Chota, Jaén y San Ignacio), el diseño consideró seleccionar aquí una muestra ppt (proporcional al número de internos de cada pabellón) de 4 pabellones para internos hombres (de un total de 9 pabellones que debería considerar) y solo 2 pabellones de internas mujeres. Finalmente, para los pabellones seleccionados de hombres se tomó un MASs de 10 internos y en el de las de mujeres se entrevistaron a todas las mujeres de los dos pabellones. Dado que los otros dominios contaban, con tan solo un pabellón, en estos se procedió a tomar un MASs de 30 reos hombres por penal y censar a las mujeres en estos penales. Tomándose la muestra correspondiente y considerándose el censo penal del 2016 solo como marco muestral,

a) Estime para el dominio de Chota la proporción de internos sentenciados y su error estándar de estimación.

b) Estime para el dominio de Cajamarca la proporción de internos sentenciados y su error estándar de estimación.

c) Dé una estimación de los otros dos parámetros de interés tanto en Cajamarca como en Chota.

d) ¿Cuál sería la estimación y el error estándar de estimación estimado para la proporción de sentenciados en toda la región de Cajamarca?

e) ¿Podría concluir, a un nivel de significación de $\alpha = 0.05$ que en esta región existe asociación entre el consumo de drogas de los internos y el tipo de delito cometido?

14. En un terreno se ha resembrado una especie de árbol y se desea estimar, entre otras cosas, la altura promedio de estos árboles pasados 5 años de iniciada la reforestación. El terreno se dividió en 50 lotes de tamaños similares, donde 30 lotes están bajo la administración de una compañía privada y 20 bajo la administración de una compañía estatal. Suponga que le brindan la información que aparece en el cuadro 5.2, donde se consignan las alturas en centímetros de un conjunto de árboles seleccionados mediante un MASs en cada uno de 5 lotes también seleccionados por un MASs dentro de cada administración, así como el número de árboles en cada lote seleccionado que mostraron ya algún signo de floración

a) Estime la altura promedio de los árboles en el terreno y su error de estándar de estimación. ¿Cuál sería la estimación de la desviación estándar de las alturas de estos árboles?

b) Estime la proporción de árboles en el terreno que muestran signos de floración y su error de estándar de estimación.

c) Estime los efectos de diseño en las estimaciones anteriores.

Administración	Núm. de árboles	Núm. de árboles con floración	Núm. de árboles muestreados	Altura de los árboles
Privada	52	30	5	32.72, 26.68, 29.42, 24.52, 34.04
Privada	56	35	6	25.43, 23.30, 18.32, 23.08, 20.68, 24.24
Privada	51	28	5	35.47, 37.18, 31.32, 29.08, 34.24
Privada	49	39	5	24.10, 27.50, 34.30, 30.88, 31.26
Privada	45	33	6	30.88, 27.78, 27.84,32.23, 28.83, 30.03
Estatal	60	26	6	16.47, 12.77, 17.28, 13.14, 15.53, 10.13
Estatal	46	25	5	17.62, 20.20, 17.73,18.21, 15.32
Estatal	50	37	5	14.86, 18.36, 14.82, 21.37, 17.09
Estatal	61	33	6	23.53, 26.66, 21.30, 22.76, 23.48, 23.26
Estatal	60	34	6	18.09, 25.29, 20.80, 22.96, 24.03, 25.23

Cuadro 5.2: Datos del diseño para el ejercicio 14

15. Se desea estimar el rendimiento medio en lectura de los estudiantes del segundo año de educación secundaria de la provincia constitucional del Callao. Considere, como marco muestral, la ECE 2019 y utilice un muestreo estratificado de conglomerados bietápico. Los estratos estarán definidos por el tipo de gestión del colegio (Estatal y No estatal). En cada estrato se considerarán conglomerados de primera etapa a los colegios y de segunda etapa a los alumnos. Para simplificar, puede suponer que dispone de un presupuesto limitado que solo alcanza para seleccionar a 30 colegios y a un máximo de 20 alumnos por colegio.

a) Obtenga una estimación del rendimiento medio en lectura de los estudiantes del segundo año de secundaria del Callao, así como su error estándar de estimación.

b) Obtenga una estimación del rendimiento medio en lectura de los estudiantes del segundo año de secundaria por cada estrato, así como sus errores estándares de estimación.

c) Si para el estrato estatal utiliza un muestreo con probabilidades proporcionales al número de estudiantes, en la primera etapa, y un MASs de estudiantes en la segunda etapa, mejoría su estimación del rendimiento medio en lectura?

d) Obtenga el número total estimado de profesores en los colegios del Callao. No tiene que hacer aquí un trabajo de campo para obtener tal información, pero sí puede utilizar, por ejemplo, la página web ESCALE del Ministerio de Educación que tiene información actualizada de todos los colegios en el país. Reporte aquí el intervalo de confianza al 95 % para este total y estime el efecto de diseño.

16. Suponga que para estimar el índice de rendimiento medio api para el 2000 en la población api estudiada en clase usted decide realizar un muestreo complejo que consiste en utilizar primero un MAE usando como variable de estratificación el tipo de colegio y tomando luego, con probabilidades proporcionales al tamaño (de la variable `enroll),` un muestreo por conglomerados de, respectivamente, 10 distritos con colegios de tipo elemental, 5 con colegios de tipo medio y 5 con colegios de tipo alto.
a) Estime el índice de rendimiento medio api del 2000 para la población de escuelas públicas de California y de sus estratos, así como el índice que debería haber obtenido una escuela en ese año para ser considerado en el tercio superior.
b) Use, bajo este diseño, un análisis de regresión para analizar si la variable `emer` afecta al índice api del 2000.

17. En el paquete survey de R tiene disponible la base de data poblacional elections, en donde se muestran la cantidad de votos que los tres candidatos Bush, Kerry y Nader alcanzaron en cada uno de los 4600 condados en su postulación para la presidencia de los Estados Unidos en el 2004.
a) Use el método de Tillé para seleccionar una muestra sin reemplazamiento de 40 condados con probabilidades proporcionales al número de votos alcanzados en estos. Adicione luego a esta base de datos ppsample las probabilidades de selección y los pesos de muestreo wt.
b) Basándose en la muestra anterior y el diseño

```
ppsr <- svydesign(id=~1,weight=~wt, data = ppsample),
```

estime el total de votos que cada uno de estos candidatos alcanzó en las elecciones del 2004. Indique qué es lo que este diseño asume.
c) Realice un pequeño estudio de simulación al replicar b) 1000 veces. Compare luego la media de los totales estimados con los reales y obtenga intervalos de confianza al 95 % para los totales poblacionales calculando la cobertura sobre los verdaderos valores.
d) Obtenga las estimaciones de Horvitz-Thompson para los totales pedidos y sus errores estándar de estimación estimados. Compare estas con las estimaciones en b).

Apéndice A

Sugerencias o respuestas a los problemas pares

Este apéndice incluye algunas sugerencias o soluciones a los problemas pares planteados en el texto. Para efectos de replicación y uniformidad, usaremos en lo posible la semilla aleatoria `set.seed(12345)`.

Capítulo 1

2. a) Si X denota el número de vales de 50 soles que Juan obtendrá y la selección es con reemplazamiento, entonces $X \sim B(4, \frac{1}{5})$. Si no hay reemplazamiento, $X \sim H(5, 1, 4)$. Así $P(X \geq 1)$ es mayor en el segundo caso, pues en R `1-dbinom(0,4,0.2)`$= 0.5904$ y `1-dhyper(0,1,4,4)`$= 0.8$.
b) Sea $(X_1, X_2, X_3, X_4, X_5) \sim Mul(4; 0{,}2, 0{,}2, 0{,}2, 0{,}2, 0{,}2)$ el vector aleatorio que denota el número de veces que ganarán, respectivamente, Juan, Pepe, Rosa, Luis y María un monto de 50 soles. Entonces, marginalizando $P(X_1 = 1, X_3 = 2) = 0.0576$. Esta no coincide con la probabilidad $P(X_1 = 3) = 0.0256$ de que Juan gane 300 soles.
c) Considerando a Rosa y Luis como una sola categoría, su distribución para el número de vales ganados entre los dos es binomial y, por tanto, la probabilidad de que ellos ganen los 4 sorteos es $(\frac{2}{5})^4 = 0.0256$.
d) El monto que Juan obtendrá es $M = 50X_1$ y su esperado es de 40 soles.

4. a) Sea (X_1, X_2, X_3) el vector aleatorio cuyas componentes denotan, respectivamente, el número de artículos con defectos de tipo A, B y sin defecto en la muestra de los 20 artículos de la producción. Por construcción, $(X_1, X_2, X_3) \sim HMul(20; 12, 8, 180)$ y la utilidad por vender estos artículos es $U = 25X_3 - 5X_1 - 10X_2$. Se pide en principio $P(U = 400) = P(25(20 - X_1 - X_2) - 5X_1 - 10X_2 = 400) = P(500 - 30X_1 - 35X_2 = 400) = P(6X_1 + 7X_2 = 20) = P(X_1 = 1, X_2 = 2, X_3 = 17)$. Esto se calcula en R por

```
choose(12,1)*choose(8,2)*choose(180,17)/choose(200,20)

## [1] 0.0587
```

b) Por otro lado, el valor esperado de U en soles es

$$E(U) = 25E(X_3) - 5E(X_1) - 10E(X_2) = 25 \times 20 \times \frac{180}{200} - 5 \times 20 \times \frac{12}{200} - 10 \times 20 \times \frac{8}{200} = 436,$$

mientras que la varianza de U es igual a

$$\begin{aligned} V(U) &= 625V(X_3) + 25V(X_1) + 100V(X_2) - 250Cov(X_3, X_1) - 500Cov(X_3, X_2) + 100Cov(X_1, X_2) \\ &= \frac{3600}{7\,960\,000}(625(180)(20) + 25(12)(188) + 100(8)(192) + 250(180)(12) \\ &\qquad + 500(180)(8) - 100(12)(8)) = 1678.07. \end{aligned}$$

Por tanto, la desviación estándar de las utilidades es de 40.96426 soles.

6. El número de personas entrevistadas en la encuesta más pequeña, que ya habían sido entrevistadas en la encuesta más grande, X, satisface $X \sim H(50, 20, 10)$. Por tanto, su valor esperado y varianza vienen dados, respectivamente, por $E(X) = 4$ y $V(X) = 1.959$.

8. a) Denotemos a X_i como la v.a. que nos dice cuántas cápsulas del medicamento genérico contiene la caja i, $i = 1, 2, 3, 4$. Naturalmente, $X_1 \sim H(24, 4, 6)$.
b) Se nos pide $P(X_3 = 4)$. Dado que la selección se hace secuencialmente, podríamos reescribir esta probabilidad con la regla del producto, como

$$\begin{aligned} P(X_3 = 4) &= P(X_3 = 4 \mid X_1 = 0, X_2 = 0)P(X_2 = 0 \mid X_1 = 0)P(X_1 = 0) \\ &= \frac{C_4^4 C_2^8}{C_6^{12}} \times \frac{C_0^4 C_6^{14}}{C_6^{18}} \times \frac{C_0^4 C_6^{20}}{C_6^{24}} = \frac{C_4^4 C_2^{20}}{C_6^{24}} = 0.001411632 \end{aligned}$$

y, por tanto, esta probabilidad es la misma si se tratara de la caja 1 o de cualquier otra caja.
c) Vimos que $X_1 \sim H(24, 4, 6)$. Consecuentemente, su distribución de probabilidades es

```
dhyper(0:4,4,20,6)

## [1] 0.28797 0.46076 0.21598 0.03388 0.00141
```

Por otro lado, la función de probabilidad de X_2 se puede hallar al condicionar sobre X_1 mediante

```
P2 <-function(x){
x1 = c(0,1,2,3,4)
sum(dhyper(x,4-x1,14+x1,6)*dhyper(x1,4,20,6))}
```

Similarmente, condicionándose a las selecciones previas, las funciones de probabilidad de X_3 y X_4 se obtienen mediante las funciones

```
P3 <-function(x){
A = matrix(0,5,5)
for(x1 in 0:4){
for(x2 in 0:(4-x1)){
ax1 = dhyper(x,4-x1-x2,8+x1+x2,6)*dhyper(x2,4-x1,14+x1,6)
A[x1+1,x2+1]=ax1*dhyper(x1,4,20,6)}}
sum(A)}
```

```
P4 <-function(x){
A = array(0,dim = c(5,5,5))
for(x1 in 0:4){
for(x2 in 0:(4-x1)){
for(x3 in 0:(4-x1-x2)){
ax2 = dhyper(x,4-x1-x2-x3, 2+x1+x2+x3,6)*dhyper(x3,4-x1-x2,8+x1+x2,6)
A[x1+1,x2+1,x3+1]=ax2*dhyper(x2,4-x1,14+x1,6)*dhyper(x1,4,20,6)}}}
sum(A)}
```

Como se comprueba con, por ejemplo, X_4

```
c(P4(0),P4(1),P4(2),P4(3),P4(4))

## [1] 0.28797 0.46076 0.21598 0.03388 0.00141
```

todas estas funciones nos brindan la misma distribución que la de X_1.
d) Como el rango del vector (X_1, X_2, X_3, X_4) son los números naturales cuya suma es 4, se tiene que

$$P_{X_1X_2X_3X_4}(x_1,x_2,x_3,x_4) = P(X_4 = x_4 \mid X_1 = x_1, X_2 = x_2, X_3 = x_3)P(X_3 = x_3 \mid X_2 = x_2, X_1 = x_1)$$

$$\times P(X_2 = x_2 \mid X_1 = x_1)P(X_1 = x_1)$$

$$= \mathbf{1}_{x_4}(x_4)\frac{C_{x_3}^{4-x_1-x_2}C_{6-x_3}^{8+x_1+x_2}}{C_6^{12}} \times \frac{C_{x_2}^{4-x_1}C_{6-x_2}^{14+x_1}}{C_6^{18}} \times \frac{C_{x_1}^{4}C_{6-x_1}^{20}}{C_6^{24}} = \frac{C_{x_1}^{6}C_{x_2}^{6}C_{x_3}^{6}C_{x_4}^{6}}{C_4^{24}}.$$

e) Sea Y la v.a. que denota el número de cajas que contienen alguna cápsula genérica. El rango de esta v.a es $R_Y = \{1, 2, 3, 4\}$ y se tiene que

$$P(Y = 1) = P(X_1 = 4)+P(X_2 = 4)+P(X_3 = 4)+P(X_4 = 4) = 4P(X_1 = 4) = 0.005646527.$$

$$P(Y=2) = C_2^4 P(X_1=2, X_2=2, X_3=0, X_4=0) + C_2^4 P(X_1=1, X_2=3, X_3=0, X_4=0)$$
$$+C_2^4 P(X_1=3, X_2=1, X_3=0, X_4=0) = 0.2625635.$$
$$P(Y=4) = P(X_1=1, X_2=1, X_3=1, X_4=1) = 0.121965$$

y, por complemento, $P(Y=3) = 0.609825$.

10. Formalmente, $(X_i, X_j, X_m, X_o) \sim Hmul(n; M_i, M_j, M_m, N - M_i - M_j - M_m)$, donde X_o denota los elementos seleccionados en la muestra que no pertenecen a las clases i, j, ni m; sin embargo, para cualquier x_1, x_2 y x_3 entero, siempre se cumple que

$$P(X_i = x_1, X_j = x_2, X_m = x_3) = P(X_i = x_1, X_j = x_2, X_m = x_3, X_o = n - \sum_{i=1}^{3} x_i)$$
$$= \frac{C_{x_1}^{M_i} C_{x_2}^{M_j} C_{x_3}^{M_m} C_{n-\sum_{i=1}^3 x_i}^{N-M_i-M_j-M_m}}{C_n^N}.$$

Decir, por tanto, que el vector aleatorio (X_i, X_j, X_m) tiene distribución hipergeométrica multivariada no es en verdad cierto, aun cuando su distribución se derive de esta última. La función de probabilidad del vector (X_i, X_j, X_m) viene dada por la expresión anterior.

12. Para encontrar el estimador, propongamos uno lineal de la forma $\hat{p} = \sum_{i=1}^6 c_i X_i \delta_i$. Condicionando, $E(\hat{p}) = \frac{1}{6}\sum_{i=1}^6 c_i \frac{n_i M_i}{N_i}$. El valor de la constante c_i que hace que este sea insesgado es, por tanto, $c_i = \frac{N_i}{n_i \bar{N}}$, donde $\bar{N} = \frac{1}{6}\sum_{i=1}^6 N_i$. El estimador insesgado de p es por tanto,

$$\hat{p} = \frac{1}{\bar{N}} \sum_{i=1}^{6} N_i \bar{p}_i \delta_i.$$

La varianza de este estimador viene, por la proposición 1.5, dada por

$$V(\hat{p}) = E\left(V(\hat{p} \mid \delta_1, \ldots, \delta_6)\right) + V\left(E(\hat{p} \mid \delta_1, \ldots, \delta_6)\right).$$

Como las muestras a tomar en cada caja se pueden asumir independientes, se tendrá que

$$V(\hat{p} \mid \delta_1, \ldots, \delta_6) = \frac{1}{\bar{N}^2} \sum_{i=1}^{6} V(\bar{p}_i) N_i^2 \delta_i^2 = \frac{1}{\bar{N}^2} \sum_{i=1}^{6} \frac{1}{n_i} \frac{M_i}{N_i} (1 - \frac{M_i}{N_i})(\frac{N_i - n_i}{N_i - 1}) N_i^2 \delta_i^2$$

y así,

$$E\left(V(\hat{p} \mid \delta_1, \ldots, \delta_6)\right) = \frac{1}{\bar{N}^2} \sum_{i=1}^{6} \frac{1}{n_i} \frac{M_i}{N_i} (1 - \frac{M_i}{N_i})(\frac{N_i - n_i}{N_i - 1}) N_i^2 \frac{1}{6} = \frac{1}{6\bar{N}^2} \sum_{i=1}^{6} \frac{M_i(N_i - M_i)(N_i - n_i)}{n_i(N_i - 1)}.$$

Por otro lado, como $E(\hat{p} \mid \delta_1, \ldots, \delta_6) = \frac{1}{\bar{N}} \sum_{i=1}^6 N_i p_i \delta_i$, se tiene que

$$V\left(E(\hat{p} \mid \delta_1, \ldots, \delta_6)\right) = \frac{1}{\bar{N}^2} \left(\sum_{i=1}^{6} N_i^2 p_i^2 V(\delta_i) + \sum_{\substack{i=1 \\ i \neq j}}^{6} \sum_{j=1}^{6} N_i N_j p_i p_j Cov(\delta_i, \delta_j) \right)$$

$$= \frac{1}{36\bar{N}^2}\left(5\sum_{i=1}^{6} N_i^2 p_i^2 - \sum_{\substack{i=1 \\ i\neq j}}^{6}\sum_{j=1}^{6} N_i N_j p_i p_j\right).$$

Por tanto, la varianza pedida viene dada por

$$V(\hat{p}) = \frac{1}{6\bar{N}^2}\left(\sum_{i=1}^{6}\frac{M_i(N_i-M_i)(N_i-n_i)}{n_i(N_i-1)} + \frac{5}{6}\sum_{i=1}^{6} N_i^2 p_i^2 - \frac{1}{6}\sum_{\substack{i=1 \\ i\neq j}}^{6}\sum_{j=1}^{6} N_i N_j p_i p_j\right).$$

14. a) Si X denota la cantidad de personas encuestadas de las tres primeras instituciones, entonces $X \sim H(M, M_1+M_2+M_3, n)$, donde $M = \sum_{i=1}^{N} M_i$. Se pide

$$P(X=n) = \frac{C_n^{M_1+M_2+M_3}}{C_n^M}.$$

b) Similiarmente, sea Y la v.a. que denota el número de personas encuestadas de la primera institución. Entonces $Y \sim H(M, M_1, n)$ y se tiene que $P(Ne=1) = P(Y=n) = \frac{C_n^{M_1}}{C_n^M}$, donde $M_1 \geq n$.
c) Como se suguiere, la v.a. Ne se puede escribir como $Ne = \sum_{i=1}^{N} \mathbf{1}_{\{X_i>0\}}$, donde $(X_1, X_2, \ldots, X_N) \sim Hmul(M; M_1, M_2, \ldots, M_N)$. Así, $E(Ne) = \sum_{i=1}^{N} E(\mathbf{1}_{\{X_i>0\}}) = \sum_{i=1}^{N} P(X_i>0) = \sum_{i=1}^{N}(1-P(X_i=0)) = \sum_{i=1}^{N}(1-\frac{C_n^{M-M_i}}{C_n^M})$.
d) La probabilidad de que la muestra esté constituida solo por participantes de las tres primeras instituciones es $P(X=16) = \frac{C_{16}^{28}}{C_{16}^{100}} = 2.26\times10^{-11}$. Por otro lado, $P(Ne=1) = \frac{C_{16}^{17}}{C_{16}^{100}} = 1.263\times10^{-17}$. Se espera, por otro lado, entrevistar a personas de aproximadamente

```
x = c(17 , 8 , 3 , 4 , 6 , 9 , 12 , 14 , 1 , 2 , 1 ,4 , 2 , 10 , 2 , 5)
round(sum(1-choose(100-x,16)/choose(100,16)))

## [1] 9
```

instituciones, donde `x` denota el vector del número de personas por institución. Finalmente, para que $Ne=2$ deberían seleccionarse cualesquiera de los siguientes conjuntos de instituciones $\{8,10\}, \{8,13\}, \{8,15\}, \{7,4\}, \{7,12\}$ y $\{14,5\}$. Por tanto, utilizándose la distribución hipergeométrica multivariada, se tendrá que

$$P(Ne=2) = P(X_8=14, X_{10}=2, X_0=0) + P(X_8=14, X_{13}=2, X_0=0)$$

$$+P(X_8=14, X_{15}=2, X_0=0) + P(X_7=12, X_4=4, X_0=0)$$

$$+P(X_7=12, X_{12}=4, X_0=0) + P(X_{14}=10, X_5=6, X_0=0),$$

donde X_0 denota el número de personas encuestadas de las otras instituciones no consideradas al interior de las probabilidades. Note que todas estas probabilidades son las mismas e iguales a $\frac{1}{C_{16}^{100}}$ y, por tanto, $P(Ne = 2) = \frac{6}{C_{16}^{100}} = 4.458 \times 10^{-18}$.

16. a) Si X denota el número de parqueos que tendrá que pagar el turista, se tiene que $X \sim H(12, 9, 4)$ y, por tanto, su valor esperado es $E(X) = \frac{4\times 9}{12} = 3$.

b) Para simular se puede usar la función `rhyper`

```
set.seed(12345)
rhyper(1,9,3,4)

## [1] 3
```

Otra manera es mediante

```
set.seed(12345)
min(which(phyper(0:4,9,3,4)>runif(1)))-1

## [1] 3
```

c) No es adecuada, pues la selección de parques en el lazo (for) es con reemplazamiento y se dice que el turista elige 4 de los 12 parques.

Capítulo 2

2. a) Note que $X \sim H(N, m, n)$. Un desarrollo de Taylor de segundo orden para $\hat{N}_1$ alrededor de la media de X, $\mu = E(X) = \frac{nm}{N}$, nos da la aproximación

$$\hat{N}_1 = \frac{nm}{\mu} - \frac{1}{\mu^2}(X - \mu) + \frac{2}{\mu^3}(X - \mu)^2.$$

Tomando el valor esperado obtendremos la primera expresión a probar. En cuanto a la varianza de $\hat{N}_1$, podríamos considerar solo el desarrollo de primer orden y obtener, tomando varianzas a esta, la aproximación

$$V(\hat{N}_1) = \frac{n^2m^2}{\mu^4}V(X) = \frac{N^4}{n^2m^2}n\frac{m}{N}(1 - \frac{m}{N})\frac{N - n}{N - 1} = \frac{N^2(N - m)(N - n)}{nm(N - 1)}.$$

b) Note que $Y \sim BN(r, p = \frac{m}{N})$, luego $E(\hat{N}_2) = \frac{m}{r}E(Y) = \frac{m}{r}\frac{r}{p} = N$. Similarmente,

$$V(\hat{N}_2) = \frac{m^2}{r^2}V(Y) = \frac{m^2}{r^2}\frac{r(1 - p)}{p^2} = \frac{N(N - m)}{r}.$$

Por otro lado,

$$E(\hat{V}(\hat{N}_2)) = \frac{m^2}{r^2(r+1)}(E(Y^2) - rE(Y)) = \frac{m^2}{r^2(r+1)}(\frac{r(1-p)}{p^2} + \frac{r^2}{p^2} - \frac{r^2}{p}) = \frac{N(N-m)}{r}.$$

Una desventaja del muestreo inverso es que el número de selecciones hasta obtener los r elementos marcados puede ser grande, lo cual hace que este sea costoso y tome mucho tiempo.

c) $\hat{N}_1 = 500$ y $\hat{N}_2 = 508$. Reemplazando en $V(\hat{N}_1)$, N por su estimación $\hat{N}_1$ y usando $\hat{V}(\hat{N}_2)$, los IC asíntoticos para N según los métodos directo e inverso vienen dados, respectivamente, por

$$[70.15546,\ 929.8445] \quad \text{y} \quad [109.5974,\ 906.4026].$$

En esta aplicación, el muestreo inverso parece ser más preciso.

4. a) Si es un estimador insesgado.

b) $V(\bar{Y}_c) = (1 - \frac{n}{N})(\frac{\sigma^2_{N-1}}{n} + \frac{2nc^2}{N-1})$.

c) No hay contradicción.

6. Recordemos que toda muestra en un MASc puede representarse por un vector $(\delta_1, \delta_2, \ldots, \delta_N)$, donde δ_i denota el número de veces que la unidad i es seleccionada. Estas v.a. toman valores en el conjunto $\{0, 1, 2, \ldots, n\}$ y satisfacen

$$\delta_1 + \delta_2 + \ldots + \delta_N = n.$$

Si identificamos ahora a cada valor entero positivo por igual número de barras verticales y mantenemos los signos +, podríamos, entonces, identificar cada muestra por una única secuencia de barras verticales y signos +. Por ejemplo, si $N = 9$ y $n = 6$ una posible muestra es que la primera unidad sea elegida 3 veces, la cuarta 2 veces y la octava una vez; esto es:

$$(3, 0, 0, 2, 0, 0, 0, 1, 0),$$

pues

$$3 + 0 + 0 + 2 + 0 + 0 + 0 + 1 + 0 = 6.$$

Así, esta muestra se representará por la secuencia

$$|||+++||++++|+$$

Consecuentemente, el número total de muestras que se podrán obtener en un MASc es igual al número de maneras que podríamos ordenar estas secuencias, donde se tienen n caracteres repetidos de tipo | y $N-1$ caracteres repetidos de tipo +. Esto es bien conocido y viene dado por la cantidad de permutaciones con elementos repetidos; es decir, por

$$\frac{(N+n-1)!}{n!(N-1)!} = C_n^{N+n-1}.$$

8. a) Usando la regla conservadora $\bar{p} = 0.5$, se tiene que $n = 86$.
b) $(X_A, X_B, X_C, X_D) \sim HMul(86; 10, 20, 8, 682)$ denota el número de fábricas que serán seleccionadas de cada consorcio y D para los que no están en un consorcio. En particular, $X_B \sim H(720, 20, 86)$ y $P(X_B > 0) = 1 - P(X_B = 0) = 0.9242674$.
c) El valor esperado es 107.75 o aproximadamente 108.

10. a) $\frac{n}{N}$.
b) $\frac{M}{N-n}$.
c) Definiendo los eventos A_i y B_i como, respectivamente, yo y mis padres seamos seleccionados en el i-ésimo día, se nos pide

$$P(A_1 \cap B_1) + P(A_2 \mid A_1^c \cap B_1)P(A_1^c \cap B_1) + P(B_2 \mid A_1 \cap B_1^c)P(A_1 \cap B_1^c)$$

$$+P(A_2 \cap B_2 \mid A_1^c \cap B_1^c)P(A_1^c \cap B_1^c).$$

Condicionando aquí las probabilidades condicionales del segundo día con respecto a la v.a. X =número de viviendas que no responden el primer día $\sim B(n, q)$, se sigue que esta probabilidad viene dada por $\frac{n}{N(N-1)}(n - 1 + 4nq + (n-1)q^2)$.
d) Podríamos agregar la v.a Y = número de viviendas que responden el segundo día. Note que $Y \mid X = x \sim B(x, 1-q)$. Así, la probabilidad de que se complete el tamaño de muestra planificado viene dada por

$$\sum_{x=0}^{n} P(Y = x \mid X = x)P(X = x) = (1 - q^2)^n$$

y la probabilidad pedida es $1 - (1 - q^2)^n$.
e) 0.3027767.

12. a) Considere la primera caracterización de S^2 y sume y reste luego $\bar{Y}$ al interior de $(Y_i - Y_j)^2$. Desarrollando el cuadrado y operando es inmediato llegar a la fórmula tradicional de S^2.
b)

$$E(S^2) = \frac{1}{2n(n-1)} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} E(Y_i - Y_j)^2 = \frac{1}{2n(n-1)} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} V(Y_i - Y_j)^2.$$

$$= \frac{1}{2n(n-1)} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} (V(Y_i) + V(Y_j) - 2Cov(Y_i, Y_j)) = \frac{1}{2n(n-1)} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} 2\sigma_{N-1}^2 = \sigma_{N-1}^2.$$

c) Basta tomar el límite cuando $N \to \infty$.

14. a) La función de probabilidad de X es $P_X(x) = C_x^N (\frac{1}{N})^5 a_x$, donde

$$a_x = 1\mathbf{1}_{x=1}(x) + 30\mathbf{1}_{x=2}(x) + 150\mathbf{1}_{x=3}(x) + 240\mathbf{1}_{x=4}(x) + 120\mathbf{1}_{x=5}(x).$$

b) Podría definir la variable dicotómica $\delta_i^* = \mathbf{1}_{\delta_i>0}$ y expresar el estimador como $\hat{\tau}^* = C\sum_{i=1}^N y_i\delta_i^*$. Sobre la base de ello, la constante que hace a este estimador insesgado es $C = \frac{1}{1-q}$, donde $q = (1-\frac{1}{N})^5$.
c) La varianza de este estimador viene dada por

$$V(\hat{\tau}_i^*) = \frac{q}{1-q}\sum_{i=1}^N y_i^2 + (\frac{(1-2q+(1-\frac{2}{N})^5}{(1-q)^2} - 1)\sum_{i=1}^N\sum_{j\neq i}^N y_i y_j.$$

d) El código

```
F = (1:15)/15
u = c(0.327, 0.894, 0.131, 0.289, 0.643)
m = NULL
for(i in 1:5) m[i] = min(which((F>u[i]) == TRUE))
m

## [1]  5 14  2  5 10
```

nos dice que la muestra está conformada por 4 personas distintas.
16. a) Sea $y_i^* = y_i\gamma_i$, entonces $E(\hat{\tau}_d) = \frac{N}{n}\sum_{i=1}^N y_i^* E(\delta_i) = \sum_{i=1}^N y_i^* = \tau_d$.
b) Dado que los N datos de y^* los podemos particionar en dos subconjuntos de tamaños N_d y $N - N_d$, donde el primero contiene los datos del dominio y el segundo son todos 0, la media μ_{*d} de estos datos es $\mu_{*d} = \frac{N_d\mu_d}{N}$ y su varianza satisface

$$\sigma_{*d}^2 = \frac{1}{N-1}((N_d-1)\sigma_d^2 + (N-N_d-1)\times 0 + N_d\mu_d^2 + (N-N_d)\times 0 - N\mu_{*d}^2)$$

$$= \frac{1}{N-1}((N_d-1)\sigma_d^2 + q_d N_d\mu_d^2).$$

c) Como $\hat{\tau}_d = N\bar{Y}_d$, donde $\bar{Y}_d$ es la media muestral en la población estadística $\mathcal{P}_{y^*}$, se tiene por la proposición 2.2 que

$$V(\hat{\tau}_d) = N^2(1-\frac{n}{N})\frac{\sigma_{*d}^2}{n},$$

siendo σ_{*d}^2 la varianza poblacional de $\mathcal{P}_{y^*}$.
d) Basta reemplazar b) en c) y considerar la fórmula del tamaño de muestra de un total sobre la población $\mathcal{P}_{y^*}$: $n = \frac{z_{1-\frac{\alpha}{2}}^2\sigma_{*d}^2 N^2}{Nz_{1-\frac{\alpha}{2}}^2\sigma_{*d}^2 + e^2}$, la cual se obtiene del de la media, redefiniendo el error.
e) Solo considerar en la fórmula anterior que $e = N_d\mu_d z_{1-\frac{\alpha}{2}}CV_0$ y aproximar de manera natural las fracciones de los tamaños en la población total y del dominio.
f) Se nos brindan las estimaciones $\hat{\mu}_d = 5100$ y $\hat{\sigma}_d = 380$, con lo cual la estimación actual del consumo total de agua en la zona será de 15 millones y 300 000 litros. Dado que desconocemos

N_d (asumiendo que en el trabajo de campo no hubo presupuesto para obtener este valor), podríamos usar la aproximación dada en d) con las estimaciones anteriores y $\bar{p}_d = 0.6$. Así, el tamaño de muestra requerido será de $n = 204$ viviendas.

18. a) El tamaño de muestra requerido se calculará sobre la base de las estimaciones del mismo índice en 1999 como:

```
N = dim(apipop)[1]
z = qnorm(0.975)
mu0 = mean(api99)
s0 = sd(api99)
e = mu0*z*0.03
n = (N*(z*s0)^2)/((z*s0)^2 + N*e^2)
(n = ceiling(n))

## [1] 49
```

El diseño y las estimaciones son

```
set.seed(12345)
muestra = apipop[sample(N,n),]
dism = svydesign(ids= ~1,fpc= rep(N,n),data = muestra)
(m = svymean(~api00,dism))

##       mean   SE
## api00  646 17.9

(svytotal(~enroll,dism,na.rm=T))

##          total     SE
## enroll 3853806 395991

(svyby(~api00, ~stype, dism, svymean))

##   stype api00   se
## E     E   650 20.4
## H     H   665 38.1
## M     M   581 66.5
```

siendo los verdaderos valores de estos parámetros los siguientes:

```
mean(api00)

## [1] 665

sum(enroll,na.rm=T)

## [1] 3811472

as.table(by(api00,stype,mean))

## stype
##   E   H   M
## 672 634 656
```

Note que el error de estimación en la estimación del api00 es $|645.65 - 664.7126| = 19.0626$, que es menor al preestablecido de 37.1558 puntos. Por otro lado, se tiene el CV estimado y el intervalo de confianza al 95 % siguientes:

```
(CV = as.numeric(100*SE(m)/coef(m)))

## [1] 2.76

confint(m)

##       2.5 % 97.5 %
## api00   611    681
```

último que contiene a la verdadera media del índice api 2000.

20. a) Basta desarrollar

$$Cov(\frac{1}{n}\sum_{i=1}^{N} x_i\delta_i, \frac{1}{n}\sum_{j=1}^{N} y_j\delta_j) = \frac{1}{n^2}(\sum_{i=1}^{N}\sum_{j=1}^{N} x_i y_j Cov(\delta_i, \delta_j)),$$

recordando que $(\delta_1, \delta_2, \ldots, \delta_N) \sim Hmul(n; 1, \ldots, 1)$.
b) Un estimador natural para esta covarianza estará dada por

$$\hat{Cov}(\bar{X}, \bar{Y}) = (1 - \frac{n}{N})\frac{S_{xy}}{n},$$

donde:

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})\delta_i$$

es la covarianza muestral entre x e y. No es difícil mostrar que este es un estimador insesgado de la covarianza anterior.

22. a) Utilizando el paquete survey

```
set.seed(12345)
(index1 = sample(100,20))

##  [1] 73 87 75 86 44 16 31 48 67 91  4 14 65  1 34 40 33 97 15 78
```

Las áreas de los rectángulos seleccionados son

```
aream=c(10,48,8,12,40,24,54,54,56,40,10,8,14,12,50,20,3,42,6,30)
areas1 = data.frame(aream)
```

y las estimaciones pedidas las calculamos mediante

```
disMASs = svydesign(id=~1,fpc = rep(100,20),data=areas1)
(m1 = svytotal(~aream, disMASs))

##       total  SE
## aream  2705 379

confint(m1,level=0.98)

##        1 % 99 %
## aream 1823 3587
```

b) Para el MASc tenemos

```
set.seed(12345)
(index2 = sample(100,20,replace=TRUE))

##  [1] 73 88 77 89 46 17 33 51 73 99  4 16 74  1 40 47 39 41 18 96

aream=c(10,36,100,18,21,56,3,49,10,60,10,24,27,12,20,8,30,45,56,6)
areas2 = data.frame(aream)
disMASc = svydesign(ids=~1,weights = rep(5,20),data=areas2)
(m2 = svytotal(~aream, disMASc))

##       total  SE
## aream  3005 549

confint(m2,level=0.98)

##        1 % 99 %
## aream 1728 4282
```

24. a) Con los datos dados creamos la base de datos TallaS.RData. Las estimaciones pedidas se obtendrán mediante el código

```
load("TallaS.RData")
disTS = svydesign(id=~1,fpc=rep(700,35),data = TallaS)
m = svymean(~Estatura,disTS)
svyvar(~Estatura,disTS)

##          variance SE
## Estatura  0.00721  0

svymean(~Sexo,disTS)

##              mean   SE
## SexoHombre 0.686 0.08
## SexoMujer  0.314 0.08
```

b) El error será

```
as.numeric(qnorm(0.975)*SE(m))

## [1] 0.0274
```

c) No sería adecuado.
d) $n = \frac{z^2_{1-\frac{\alpha}{2}}\sigma^2 N}{z^2_{1-\frac{\alpha}{2}}\sigma^2 + e^2 N} = \frac{\sigma^2/\mu^2}{\sigma^2/((N\mu^2)+CV_0^2)}$. Estimando los parámetros μ y σ^2 con los datos de la muestra y fijándose $CV_0 = 0.005$, obtendremos que $n = 84$.

26. a) 0.024451.
b) EL IC contiene a 0.5, por lo cual no podría asegurarse que el candidato opositor vaya a ganar las elecciones.

28. a) El código en R sería

```
set.seed(12345)
N = dim(apipop)[1]
n = 500
index = sample(N,n)
sample = apipop[index,]
disMASs = svydesign(id=~1,fpc=rep(N,n),data = sample)
means = svymean(~api00+api99,disMASs)
(contr = svycontrast(means,c(api00=1,api99=-1)))
```

```
##          contrast   SE
## contrast     30.5 1.23
```

b) Se nos pide

```
confint(contr)

##          2.5 % 97.5 %
## contrast  28.1   32.9
```

c) Considere la variable $d = y - x$, que es la diferencia entre los índices api para el 2000 y 1999. El TLC para el esquema MASs de la sección 2.2 permitirá, asumiendo muestras y poblaciones grandes, construir el siguiente IC al $100(1-\alpha)\,\%$ para la diferencia de medias del índice api entre el 2000 y 1999:

$$IC = [\bar{D} - z_{1-\frac{\alpha}{2}} SE(\bar{D}) \ , \ \bar{Y} + z_{1-\frac{\alpha}{2}} SE(\bar{D}) \ ],$$

donde el error estándar de estimación de la diferencia de medias $SE(\bar{D}) = \sqrt{V(\bar{D})} = \sqrt{V(\bar{X}) + V(\bar{Y}) - 2Cov(\bar{X}, \bar{Y})}$ podría estimarse, según la proposición 2.2 y el ejercicio 20, por

$$\hat{SE}(\bar{D}) = \sqrt{\frac{1}{n}(1 - \frac{n}{N})(S_x^2 + S_y^2 - 2S_{xy})}$$

Realizando los cálculos, obtendremos

```
Dbar = mean(sample$api00 - sample$api99)
Sx2 = var(sample$api99)
Sy2 = var(sample$api00)
Sxy = cov(sample$api99,sample$api00)
e = 1.96*sqrt((1 - n/N)/n)*sqrt(Sx2+Sy2-2*Sxy)
c(Dbar-e,Dbar+e)

## [1] 28.1 32.9
```

valores que son prácticamente iguales a los obtenidos con el paquete survey.

Capítulo 3

2. a) Un estimador insesgado natural de μ_D es $\hat{\mu}_D = \bar{Y}_1 - \bar{Y}_2$ y el de su error estándar de estimación es

$$\hat{V}(\hat{\mu}_D) = (1 - \frac{n_1}{N})\frac{S_1^2}{n_1} + (1 - \frac{n_2}{N})\frac{S_2^2}{n_2}.$$

b) Bastará resolver

$$\begin{array}{cc} \text{mín}_{n_1,n_2} & (1-\frac{n_1}{N})\frac{\sigma_1^2}{n_1}+(1-\frac{n_2}{N})\frac{\sigma_2^2}{n_1}, \\ \text{s.a} & n_1+n_2=n \end{array}$$

cuya solución es $n_1=\frac{\sigma_1^2 n}{\sigma_1^2+\sigma_2^2}$ y $n=n-n_1=\frac{\sigma_2^2 n}{\sigma_1^2+\sigma_2^2}$.
c) $n_1=124$ y $n_2=176$.

8. Puesto que en un MAE $\bar{Y}$ y $\hat{V}(\bar{Y})=\sum_{h=1}^{H}(\frac{N_h}{N})^2(1-\frac{n_h}{N_h})\frac{S_h^2}{n_h}$ son, respectivamente, estimadores insesgados de los parámetros μ y $V(\bar{Y})$ de la población estadística $\mathcal{P}_y$ de una variable y, se tiene que

$$E(\hat{V}_{MASs}(\bar{Y}))=\frac{(N-n)}{n(N-1)}(\frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{N_h}y_{hi}^2E(\delta_{hi})-E(\bar{Y}^2)+V(\bar{Y}))$$

$$=\frac{(N-n)}{n(N-1)}(\frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}y_{hi}^2-E(\bar{Y})^2)=\frac{(N-n)}{n(N-1)}(\frac{1}{N}(\sigma_{N-1}^2(N-1)+N\mu^2)-\mu^2)=(1-\frac{n}{N})\frac{\sigma_{N-1}^2}{n}.$$

6. Utilizando la semilla aleatoria `set.seed(12345)` y una asignación proporcional obtendremos en R una estimación del total de 2935 unidades, con un error de estimación estimado de 176.71. El IC pedido será [2523.914 , 3346.086]. Si bien la estimación con el MASc es, por azar, ligeramente más cercana al verdadero valor, hay que apreciar que el MAE nos brinda estimaciones más confiables que la del MAS, pues su error estándar de estimación estimado es mucho menor.

8. a) Si $\mathbf{X}=(X_1,X_2,\ldots,X_H)\sim Hmul(n;N_1,N_2,\ldots,N_H)$ es el vector aleatorio que denota los tamaños de muestra en los pos(estratos), entonces la distribución marginal de X_h es hipergeométrica y se cumple que $E(X_h)=n\frac{N_h}{N}$ y $V(X_h)=n\frac{N_h}{N}(1-\frac{N_h}{N})\frac{N-n}{N-1}$. Así,

$$E(\bar{Y})=\sum_{h=1}^{H}\frac{N_h}{N}E(\bar{Y}_h)=\sum_{h=1}^{H}\frac{N_h}{N}E(E(\bar{Y}_h|X_h))$$

y

$$E(E(\bar{Y}_h|X_h))=\sum_{n_h}E(\bar{Y}_h|X_h=n_h)P(X_h=n_h)=\sum_{n_h}\mu_hP(X_h=n_h)=\mu_h,$$

donde la suma va sobre todos los posibles valores que puede tomar la distribución hipergeométrica marginal de X_h y la última igualdad se da por ser $\bar{Y}_h$ un estimador condicionalmente insesgado de μ_h.
b) Puesto que

$$V(\bar{Y}|\mathbf{X})=\sum_{h=1}^{H}(\frac{N_h}{N})^2(1-\frac{X_h}{N_h})\frac{\sigma_h^2}{X_h}=\sum_{h=1}^{H}(\frac{N_h}{N})^2\sigma_h^2(\frac{1}{X_h}-\frac{1}{N_h}),$$

la varianza (no condicionada) de $\bar{Y}$ puede obtenerse mediante

$$V(\bar{Y})=E(V(\bar{Y}|\mathbf{X}))+V(E(\bar{Y}|\mathbf{X}))=\sum_{h=1}^{H}(\frac{N_h}{N})^2\sigma_h^2(E(\frac{1}{X_h})-\frac{1}{N_h}).$$

c) $\hat{N}_h = \frac{n}{N} X_h$.
d) Como $E(\frac{1}{X_h})$ no tiene expresión conocida, podríamos considerar la expansión de Taylor de la función $f(x) = \frac{1}{x}$ evaluada en X_h hasta la segunda derivada alrededor de $E(X_h)$ y tomar esperados para obtener la aproximación:

$$E(\frac{1}{X_h}) \cong \frac{1}{E(X_h)} + \frac{1}{E(X_h)^3} V(X_h) = \frac{N}{nN_h} + (\frac{N}{nN_h})^2 (1 - \frac{N_h}{N}) \frac{N-n}{N-1}.$$

Reemplazando

$$V(\bar{Y}) \cong \frac{N-n}{nN} \sum_{h=1}^{H} (\frac{N_h}{N}) \sigma_h^2 + \frac{1}{n^2} (\frac{N-n}{N-1}) \sum_{h=1}^{H} (\frac{N-N_h}{N}) \sigma_h^2.$$

e) Los estimadores incondicionales y condicionales se calculan respectivamente con

```
set.seed(12345)
N = dim(apipop)[1]
n = 100
index1 = sample(N,n)
sam = apipop[index1,]
FreqNh = table(awards=apipop$awards)
Nh = as.vector(FreqNh)
Sh = as.vector(by(sam$api00,sam$awards,sd))
Vc = ((N-n)/(n*N))*sum((Nh/N)*Sh^2)
Vi = Vc + ((N-n)/((N-1)*n^2))*sum((N-Nh)*Sh^2/N)
c(Vi,Vc)

## [1] 159 158
```

Cabe comentar que el paquete survey no utiliza estos estimadores, sino uno propuesto por Valliant (1993) basado en residuales. Este nos provee de la siguiente estimación de la varianza de la media bajo pos(estratificación):

```
disMASs = svydesign(ids=~1,fpc=rep(N,n),data = sam)
dispost = postStratify(disMASs,~awards,FreqNh)
m = svymean(~api00,dispost)
SE(m)^2

##       api00
## api00   165
```

10. a) La probabilidad es 0.1328151.

b) Dado que en la muestra piloto se tiene información estimada de las proporciones, sugeriríamos una asignación de Neyman, lo que nos da $n = 336$.

12. Con la asignación de Neyman, los tamaños de muestra por estrato de obreros, técnicos y administradores serían, respectivamente, 45, 26 y 5; mientras que con la proporcional, 41, 29 y 9.

14. a) En este caso, la variable sexo define dos dominios de estudio, por lo cual obtendremos lo pedido mediante

```
load("ece19Am.RData")
dis19MAE = svydesign(id=~1,strata=~Estrato,fpc=~fpc,data=me19Am)
svyby(~M500_M,~sexo,dis19MAE,svymean)

##          sexo M500_M   se
## Hombre Hombre    534 4.29
## Mujer   Mujer    521 3.97
```

b) Podríamos tomar en primer lugar el estrato estatal y considerar que en esta población se tiene una estratificación por área. Luego podríamos obtener la media $\bar{Y}_{mE}$ del dominio de mujeres bajo este diseño parcial. De manera similar, obtendríamos para el diseño parcial estratificado no estatal la media $\bar{Y}_{mNE}$ del dominio de mujeres. Puesto que las muestras son independientes, la media μ_D de las diferencias en rendimiento para Matemáticas entre los dominios de estudiantes mujeres de colegios estatales y no estatales se podría estimar con su correspondiente media muestral $\bar{D} = \bar{Y}_{mE} - \bar{Y}_{mNE}$ y un IC aproximado para μ_D tendrá la forma $\bar{D} \pm z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\bar{Y}_{mE}) + \hat{V}(\bar{Y}_{mNE})}$, donde las varianzas se pueden estimar a partir de (3.5). Dependiendo de si este contiene el valor 0 o no, podremos afirmar al $100(1-\alpha)\,\%$ si existen o no diferencias significativas.
c) La misma estrategia funciona para los hombres.

16. a) Definiendo las bases de datos y calculando los tamaños de muestra:

```
load("ece19.RData")
ece19Cz = ece19[ece19$Departamento==levels(ece19$Departamento)[8],]
ece19Cz$Estrato=interaction(ece19Cz$area,ece19Cz$gestion2)
save(ece19Cz,file='ece19Cz.RData')
load("ece18.RData") # Base de datos 2018
ece18Cz = ece18[ece18$Departamento==levels(ece18$Departamento)[8],]
ece18Cz$Estrato=interaction(ece18Cz$Area,ece18Cz$Gestion2)
ece18Cz = ece18Cz[order(ece18Cz$Estrato),]
sigmah_e = by(ece18Cz$M500_M,ece18Cz$Estrato,sd,na.rm=T)
sigmah_e = as.vector(sigmah_e)
```

```
Nh = as.vector(table(ece19Cz$Estrato))
ah = Nh*sigmah_e/sum(Nh*sigmah_e)
d = dim(ece19Cz)[1]*5/qnorm(0.975)
n = sum(((Nh*sigmah_e)^2)/ah)/(d^2 + sum(Nh*sigmah_e^2))
(n = ceiling(n))

## [1] 929

(nh = round(ah*n))

## [1] 649 154 124   2
```

b) Las estimaciones pedidas estarán dadas por

```
library(sampling)
set.seed(12345)
mCz = strata(ece19Cz,c("Estrato"),size=nh,method="srswor")
me19Cz = getdata(ece19Cz,mCz)
disMAECz = svydesign(ids=~1,strata=~Estrato,fpc = rep(Nh,nh),data=me19Cz)
(meanECz = svymean(~M500_M,disMAECz,deff=T,na.rm=T))

##          mean     SE DEff
## M500_M 566.77   2.89 0.89
```

c) Puesto que las muestras en los dominios de Cusco y Amazonas son independientes, el IC al 95 % pedido viene dado por

```
SE = sqrt(SE(meanECz)^2 + SE(meanEAm)^2)
LI = coef(meanECz)-coef(meanEAm) - qnorm(0.975)*SE
LD = coef(meanECz)-coef(meanEAm) + qnorm(0.975)*SE
c(LI,LD)

## [1] 31 47
```

lo cual revela que el rendimiento medio en Matemáticas de los alumnos del Cusco es significativamente mayor que el de Amazonas.

18. a) Las estimaciones serían $\hat{\sigma}_E = 166.9314, \hat{\sigma}_H = 389.8619$ y $\hat{\sigma}_M = 353.2799$.
b) $n_E = 627, n_H = 250, n_M = 306$.
c) El número de matriculados se estima en 3 846 275 estudiantes con un IC al 95 % de [3 755 051 , 3 937 499].
d) La estimación sería de 0.68027 con un error de estimación estimado de 0.0123.

20. La base de datos se obtuvo de la página web de Amazon (usando el paquete rvest de R). Naturalmente, desde esa fecha hasta la presente deben haber cambios. La muestra y las estadísticas pedidas se obtendrán con el código siguiente:

```
library(sampling)
library(survey)
library(stringr)
load("AmazonStat.RData")
AmazonStat = AmazonStat[order(AmazonStat$tipos),]
Nh = table(AmazonStat$tipos)
nh = round(70*Nh/sum(Nh))
set.seed(12345)
me=strata(AmazonStat,c("tipos"),size=nh,method="srswor")
meAmazon = getdata(AmazonStat,me)
disme = svydesign(id=~1,strata=~tipos,fpc=~rep(Nh,nh),data=meAmazon)
(mprecios = svymean(~precios,disme))

##         mean   SE
## precios 42.3 3.78

(mstar = svymean(~starsf,disme,na.rm=T))

##        mean   SE
## starsf 4.86 0.57

aux = unlist(lapply(meAmazon$fechas, str_sub, 9,12))
(mp2017 = svymean((aux=="2017"),disme))

##        mean   SE
## [1,] 0.0857 0.03
```

La réplica de este ejercicio para el año actual queda como ejercicio.

Capítulo 4

2. Las estimaciones del total, junto con sus margenes de error al 95 %, se muestran en la tabla siguiente:

Diseño	Total estimado	Margen de error
MASc	15	16.05559
MASs	15	14.36056
MAE	10	11.68697
Sistemático	30	
Conglomerados bietápico	25	8.765225

La varianza del muestreo para la proporción por conglomerados es 0, ya que la proporción de hallazgos en los dos conglomerados resultó ser la misma, bajo la semilla dada.

4. a) Las ventas medias (utilizando un estimador de razón, que no es insesgado) en el área se estiman en $97.973 \simeq 94$ cajas por semana, con un margen de error de 63.67844 cajas semanales.
b) Si se tiene información para estimar el número total de cajas del producto A vendidas en todos los supermercados del área durante la semana. El total estimado y su error de estimacion serían, respectivamente, de 1959.46 y 223.43 cajas.

6. a) El número total de residentes jubilados se estima en 3900 con un error estándar de estimación de 635.96.
b) El número promedio de residentes jubilados por caso se estima en 0.98113 con un error estándar de estimación de 0.1127.
c) Sí se puede estimar mediante

$$\hat{\mu_\tau} = \frac{1}{4}\sum_{i=1}^{300} M_i \bar{Y}_i \delta_i,$$

donde M_i denota el número de casas en la manzana i e $\bar{Y}_i$ es la media muestral del número de jubilados por casa en la manzana i. Reemplazando, obtendremos una estimación de 13 jubilados promedio por manzana, con un error estándar de estimación estimado de 0.9.

8. Procedamos primero a demostrar el insesgamiento de los estimadores de la varianza del estimador de Horvitz-Thompson.

$$E(\hat{V}_{HT}(\hat{\tau}_{HT})) = E(E(\hat{V}_{HT}(\hat{\tau}_{HT}) \mid \delta_1, \ldots, \delta_N))$$

$$= E(\sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i^2} E(\hat{\tau_i}^2)\delta_i + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} (\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}})\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}\delta_i\delta_j + \sum_{i=1}^{N}\frac{V(\hat{\tau_i})}{\pi_i}\delta_i)$$

$$= \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i^2}(V(\hat{\tau_i}) + \tau_i^2)\pi_i + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} (\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}})\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j}\pi_{ij} + \sum_{i=1}^{N}\frac{V(\hat{\tau_i})}{\pi_i}\pi_i$$

$$= \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i}(V(\hat{\tau_i}) + \tau_i^2) + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} (\pi_{ij}-\pi_i\pi_j)\frac{\tau_i}{\pi_i}\frac{\tau_j}{\pi_j} + \sum_{i=1}^{N} V(\hat{\tau_i}) = V(\hat{\tau}_{HT}).$$

De manera similar,

$$E(\hat{V}_{SYG}(\hat{\tau}_{HT})) = E(E(\hat{V}_{SYG}(\hat{\tau}_{HT}) \mid \delta_1, \ldots, \delta_N))$$

$$= \sum_{i=1}^{N}\sum_{j>i}^{N}(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}})(\frac{V(\hat{\tau}_i)}{\pi_i^2} + \frac{V(\hat{\tau}_j)}{\pi_j^2} + (\frac{\tau_i}{\pi_i} - \frac{\tau_j}{\pi_j})^2)\pi_{ij} + \sum_{i=1}^{N} V(\hat{\tau}_i)$$

$$= \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij})(\frac{\tau_i}{\pi_i} - \frac{\tau_j}{\pi_j})^2 + \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij})(\frac{V(\hat{\tau}_i)}{\pi_i^2} + \frac{V(\hat{\tau}_j)}{\pi_j^2}) + \sum_{i=1}^{N} V(\hat{\tau}_i).$$

El segundo término a la derecha en esta expresión, que llamaremos x, resulta ser igual a

$$x = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=i}^{N}(\pi_i\pi_j - \pi_{ij})(\frac{V(\hat{\tau}_i)}{\pi_i^2} + \frac{V(\hat{\tau}_j)}{\pi_j^2}) - \sum_{i=1}^{N} V(\hat{\tau}_i).$$

Más aún, por la proposición 4.1 se tiene que

$$a = n\sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i} - (n-1)\sum_{i=1}^{N}\frac{V(\hat{\tau}_i)}{\pi_i} - \sum_{i=1}^{N} V(\hat{\tau}_i),$$

término que reemplazándose arriba en la expresión nos lleva a la ecuación dada en (5.6).

10. a) Sea π_{ij} la probabilidad conjunta de que se seleccionen en la muestra a las personas i y j. Dado que la probabilidad de inclusión π_i satisface $\pi_i = \sum_{j\neq i}\pi_{ij}$, se tendrá que

$$\begin{array}{rcccc} \pi_1 & = & 0.2 + 0.1 + 0.1 & = & 0.4 \\ \pi_2 & = & 0.2 + 0.3 + 0.15 & = & 0.65 \\ \pi_3 & = & 0.1 + 0.3 + 0.15 & = & 0.55 \\ \pi_4 & = & 0.1 + 0.15 + 0.15 & = & 0.4 \end{array}$$

b) La tabla siguiente nos muestra todas las posibles muestras de tamaño 2, así como sus probabilidades conjuntas y acumuladas:

Muestra	π_{ij}	Π_{ij}
{1,2}	0.2	0.2
{1,3}	0.1	0.3
{1,4}	0.1	0.4
{2,3}	0.3	0.7
{2,4}	0.15	0.85
{3,4}	0.15	1

En base a

```
set.seed(12345)
> runif(1)
[1] 0.7209039
```

la muestra estará conformada por las personas 2 y 4. Con ellos obtenemos la estimación (de Horvitz-Thompson) de

$$\frac{1}{0.4}+\frac{4}{0.4}=12.5;$$

es decir, de entre 12 y 13 hermanos. El error estándar de estimación estimado para este total es de 7.62 y 7.25, respectivamente, para los estimadores de Horvitz-Thompson y de Sen-Yates-Grundy.

12. Como $\boldsymbol{\delta}=(\delta_1,\delta_2,\ldots,\delta_N)\sim Mul(n;\psi_1,\psi_2,\ldots,\psi_N)$, se tiene que

$$E(\hat{\tau}_\psi)=\frac{1}{n}\sum_{i=1}^{N}E(E(\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}}{\psi_i}\mid\boldsymbol{\delta}))=\frac{1}{n}\sum_{i=1}^{N}E(\sum_{j=1}^{\delta_i}\frac{\tau_i}{\psi_i})=\frac{1}{n}\sum_{i=1}^{N}E(\delta_i)\frac{\tau_i}{\psi_i}=\frac{1}{n}\sum_{i=1}^{N}n\tau_i=\tau.$$

Por otro lado,

$$V(\hat{\tau}_\psi)=V(E(\hat{\tau}_\psi\mid\boldsymbol{\delta}))+E(V(\hat{\tau}_\psi\mid\boldsymbol{\delta}))=\frac{1}{n^2}V(\sum_{i=1}^{N}\delta_i\frac{\tau_i}{\psi_i})+\frac{1}{n^2}\sum_{i=1}^{N}E(\delta_i)\frac{V(\hat{\tau}_{ij})}{\psi_i^2}$$

$$=\frac{1}{n^2}\sum_{i=1}^{N}(\frac{\tau_i}{\psi_i})^2V(\delta_i)+\frac{1}{n^2}\sum_{\substack{i=1\\i\neq j}}^{N}\sum_{j=1}^{N}\frac{\tau_i}{\psi_i}\frac{\tau_j}{\psi_j}Cov(\delta_i,\delta_j)+\frac{1}{n}\sum_{i=1}^{N}\frac{V(\hat{\tau}_{ij})}{\psi_i}$$

$$=\frac{1}{n}(\sum_{i=1}^{N}\frac{\tau_i^2}{\psi_i}(1-\psi_i)+\sum_{i=1}^{N}\tau_i^2-(\sum_{i=1}^{N}\tau_i)^2)+\frac{1}{n}\sum_{i=1}^{N}\frac{V(\hat{\tau}_{ij})}{\psi_i}=\frac{1}{n}\sum_{i=1}^{N}\psi_i(\frac{\tau_i}{\psi_i}-\tau)^2+\frac{1}{n}\sum_{i=1}^{N}\frac{V(\hat{\tau}_{ij})}{\psi_i}.$$

Finalmente, no es difícil ver que el estimador de la varianza puede escribirse como

$$\hat{V}(\tau_\psi))=\frac{1}{n(n-1)}(\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}^2}{\psi_i^2}-n\hat{\tau}_\psi^2).$$

Así, condicionándose, la esperanza de este estimador viene dada por

$$\frac{1}{n(n-1)}(\sum_{i=1}^{N}E(\delta_i)\frac{E(\hat{\tau}_{ij}^2)}{\psi_i^2}-nV(\hat{\tau}_\psi)-nE(\hat{\tau}_\psi)^2)=\frac{1}{n(n-1)}(\sum_{i=1}^{N}n\psi_i\frac{V(\hat{\tau}_{ij})+\tau_i^2}{\psi_i^2}-nV(\hat{\tau}_\psi)-n\tau^2)$$

$$=\frac{1}{n-1}(nV(\hat{\tau}_\psi)-V(\hat{\tau}_\psi))=V(\hat{\tau}_\psi).$$

14. a) El error estándar de estimación estimado es de 0.5664.
b) La estimación de μ es 5.1.
c) La desviación estándar es 0.7248 y su estimación es 0.5818.
d) La media se estima en 5.917 y la proporción en 0.667.
e) Se distribuiría en 3 por cada zona.
f) En ambos casos la estimación sería de 5.8.

g) Podríamos obtener los efectos de diseño, donde claramente el MAE resulta ser más eficiente.

h) Estas cooperativas serán seleccionadas con probabilidad 0.0783.

i) La estimación de μ será de 5.981.

16. La estimación del número medio de personas por auto será de 4.1625 con un error estándar de estimación estimado de 0.6771.

18. a) Los distritos seleccionados serían el tercero, séptimo, décimo primero y décimo cuarto. Si se evalúan las probabilidades de inclusión de segundo orden, varias de estas asociadas a los distritos seleccionados son 0. Por lo tanto, si bien será posible estimar la proporción de colegios unidocentes pedida en aproximadamente 0.22, no será posible obtener la estimación de Horvitz-Thompson de su error estándar de estimación.

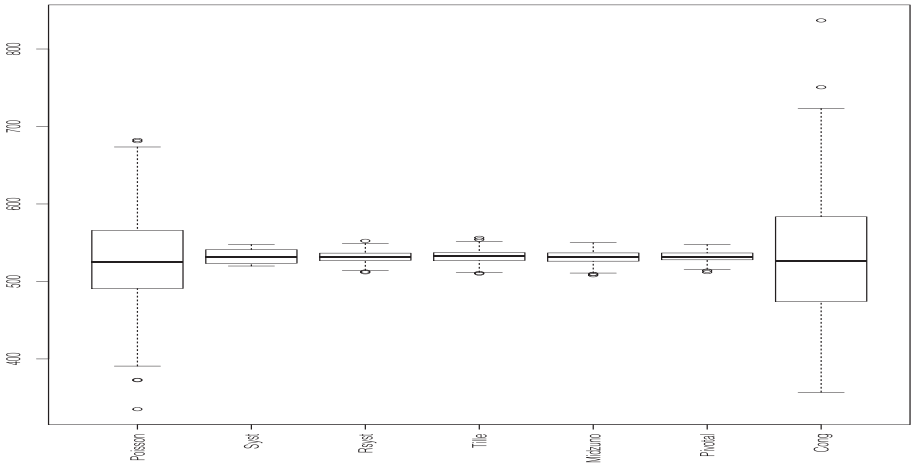
b) Ordinalmente, los distritos seleccionados bajo un muestreo por conglomerados de una etapa (con semilla aleatoria 12345) serían el 10, 11, 13 y 15; mientras que, usando el esquema de Sampdforf, obtuvimos los distritos 1, 6, 13 y 15.

c) La proporción estimada de colegios unidocentes en la región y su margen de error estimado al 95 % se muestran en la tabla siguiente:

Esquema	Proporción estimada	Margen de error
Conglomerados	0.28147	0.0675
Sampdford ppt	0.21536	0.0563

Como se observa, el muestreo ppt resultó ser más preciso, lo cual que se puede también comprobar aquí, ya que la verdadera proporción poblacional es de 0.213.

20. En este ejercicio es necesario crear una base de datos agregada de colegios con las sumas de los rendimientos evaluados. Puesto que estos últimos contienen casos perdidos, los imputaremos por su valor medio. Hecho esto, el boxplot debería quedar (ello, dependerá de las simulaciones) como por ejemplo.



Como se aprecia, todos los planes ppt, con excepción del de Poisson, muestran claramente una mayor precisión en las estimaciones del rendimiento medio en Matemáticas que el del muestreo aleatorio por conglomerados de una etapa.

22. a) La probabilidad de selección de la parcela es $\frac{6}{16}$.
b) Dado que desconocemos el número de árboles con infección avanzada por parcela no seleccionada en el área 1, utilizaremos un estimador de razón. La estimación correspondiente es de 0.2184.
c) La proporción se estima en 0.2.
d) El código en R para la estimación del caso es el siguiente:

```
set.seed(12345)
areas = c(400,580,674,920,180,300,380, 555,990,602,508,210,350,678,440,735)
num = c(16,21,18,24,24,23,25,51,42,19,11,10,36,21,37,12)
pik = inclusionprobabilities(areas,8)
m = UPsampford(pik)
HTestimator(num[m==1],pik[m==1])

##      [,1]
## [1,]  436

pik2 = UPsampfordpi2(pik)
sqrt(varHT(num[m==1],pik2[m==1,m==1],1))

## [1] 51.1
```

Capítulo 5

2. a) Puesto que el muestreo de containers es con reemplazamiento, se tiene que $X =$ número de veces que un container cualesquiera es seleccionado $\sim B(4, \frac{1}{60})$. Se nos pide, por lo tanto, $P(X > 0) = 1 - P(X = 0) = 1 - (\frac{59}{60})^4$.
b) Los pesos de las cajas para cada uno de los 4 containers seleccionados están contenidos en el vector `w` siguiente:

```
probd = (1-(59/60)^4)*3/c(100,80,114,93)
(w = 1/probd)

## [1] 513 410 584 477
```

c) El peso promedio estimado será de 11.65167 kilogramos.
d) El error estándar de estimación estimado para c) sería de 0.5078 kilogramos.
e) El cuantil 0.75 es 12.39969 kilogramos. El código sin el uso del paquete survey sería

```
Pesos = c(10.3,12.2,9.8,11.2,13.1,9.9,8.95,15.3,14.4,11.6,10.53,11.8)
Cong = rep(1:4,each=3)
Dat = data.frame(Pesos,Cong,w = rep(w,each=3))
Dat$Phat = Dat$w/sum(Dat$w)
Dat = Dat[order(Dat$Pesos),]
Dat$Fhat = cumsum(Dat$Phat)
q = 0.75
index = min(which((Dat$Fhat > q)==TRUE))
y1 = Dat$Pesos[index-1]
y2 = Dat$Pesos[index]
(qhat=y1+((q-Dat$Fhat[index-1])/(Dat$Fhat[index]-Dat$Fhat[index-1]))*(y2-y1))

## [1] 12.4
```

4. a) Empecemos considerando solo a las personas con 18 años o más

```
load("/Users/lucho/Documents/TextoMuestreo2019/Bases_de_Datos/nhis.RData")
nhis18 = nhis.large[nhis.large$age.grp>1,]
nhis18$resp = 1-as.numeric(is.na(nhis18$inc.grp))
prop.table(table(nhis18$resp))

##
##     0     1
## 0.105 0.895
```

Como se aprecia, tenemos aquí un poco más del 10 % de no respuestas a la pregunta sobre los ingresos. Para estimar las probabilidades de no respuesta utilizaremos una regresión logística con las siguientes potenciales variables predictoras:

- age.grp: Edad del adulto recodificada en 4 grupos.
- hisp: Etnicidad hispana (1 = Hispanos, 2 = No hispanos blancos, 3 = No hispanos negros 4 = Otros grupos raciales no hispanos).
- sex: Sexo (1 = Hombre, 0 = Mujer).
- race: Raza (1 = Blanca, 2 = Negra, 3 = Otra).

No se incluyeron más variables, pues estas son las únicas sin datos perdidos en la base de datos. El siguiente sería el código para el análisis de regresión logística con esta muestra compleja:

```
nhis.dsgn = svydesign(ids=~psu, strata=~stratum, data = nhis18, nest=T,
weights=~svywt)
glm.logitc = svyglm(resp~as.factor(age.grp) + as.factor(hisp)
+ as.factor(sex) + as.factor(race),
family = binomial(link="logit"),design = nhis.dsgn)

## Warning: non-integer #successes in a binomial glm!

lpredc = glm.logitc$linear.predictors
probpc = exp(lpredc)/(1 + exp(lpredc))
r = summary(glm.logitc)
```

Definamos ahora, mediante cuantiles, 5 grupos para las probabilidades de no respuesta. Las probabilidades pedidas para estos 5 grupos serán

```
q = quantile(probpc,seq(0,1,0.2))
p.class = cut(probpc,breaks=q,include.lowest=T)
phi = by(data=probpc,p.class,mean)
phi

## p.class: [0.805,0.871]
## [1] 0.857
## ------------------------------------------------------------
## p.class: (0.871,0.9]
## [1] 0.889
```

```
## ------------------------------------------------------
## p.class: (0.9,0.904]
## [1] 0.902
## ------------------------------------------------------
## p.class: (0.904,0.913]
## [1] 0.909
## ------------------------------------------------------
## p.class: (0.913,0.97]
## [1] 0.918
```

b) Para estimar la distribución étarea, por los métodos de estimación de varianza vistos en el curso, podríamos apelar al siguiente código:

```
# Definición del diseño base
nhis.dis = svydesign(id=~psu, strata=~stratum,
nest=T,data=nhis.large, weights=~svywt)
#Estimación por el método de linealización
a1 = svymean(~factor(age.grp),deff=T,design=nhis.dis)
names = c("<=18","18-24","25-44","45-64","65+")
# Estimación por los métodos de remuestreo
BRR.dis = as.svrepdesign(design=nhis.dis,type="BRR")
a2 = svymean(~factor(age.grp),deff=T,design=BRR.dis)
jkn.dis = as.svrepdesign(design=nhis.dis,type="JKn")
a3 = svymean(~factor(age.grp),deff=T,design=jkn.dis)
boot.dis=as.svrepdesign(design=nhis.dis,type="subbootstrap",replicates=1000)
a4 = svymean(~factor(age.grp),deff=T,design=jkn.dis)
list(a1,a2,a3,a4)

## [[1]]
##                       mean      SE DEff
## factor(age.grp)1 0.25309 0.00371 1.57
## factor(age.grp)2 0.10055 0.00403 3.87
## factor(age.grp)3 0.28487 0.00371 1.46
## factor(age.grp)4 0.23968 0.00420 2.09
## factor(age.grp)5 0.12181 0.00402 3.27
##
## [[2]]
##                       mean      SE DEff
## factor(age.grp)1 0.25309 0.00371 1.58
```

```
## factor(age.grp)2 0.10055 0.00406 3.93
## factor(age.grp)3 0.28487 0.00373 1.48
## factor(age.grp)4 0.23968 0.00420 2.09
## factor(age.grp)5 0.12181 0.00404 3.30
##
## [[3]]
##                       mean      SE DEff
## factor(age.grp)1 0.25309 0.00371 1.57
## factor(age.grp)2 0.10055 0.00403 3.87
## factor(age.grp)3 0.28487 0.00372 1.46
## factor(age.grp)4 0.23968 0.00420 2.09
## factor(age.grp)5 0.12181 0.00402 3.27
##
## [[4]]
##                       mean      SE DEff
## factor(age.grp)1 0.25309 0.00371 1.57
## factor(age.grp)2 0.10055 0.00403 3.87
## factor(age.grp)3 0.28487 0.00372 1.46
## factor(age.grp)4 0.23968 0.00420 2.09
## factor(age.grp)5 0.12181 0.00402 3.27
```

6. $n = 361$.

8. a) La estimación de la proporción de celulares vendidos con rebaja estará dada por

```
stock = c( 55, 45, 10, 12, 10, 120, 18, 20, 35, 45, 10, 36, 30, 27, 15, 50)
pik = inclusionprobabilities(stock,4)
w = 1/pik
select = c(1,6,9,13)
sum(c(17,35,6,13)*w[select])/sum(c(22,60,18,19)*w[select])

## [1] 0.589
```

y la estimación del monto total por ventas del celular YTRON será

```
sum(c(15395,44230,13440,13470)*w[select])

## [1] 199261
```

10. a) El código siguiente nos da la estimación pedida y su error estándar de estimación estimado:

```
dstrat<-svydesign(id=~1,strata=~stype, fpc = ~fpc,data=apistrat)
svytotal(~api.stu,dstrat)

##           total    SE
## api.stu 3086009 99477
```

b) Bastará convertir el diseño dstrat según

```
dJKn = as.svrepdesign(design=dstrat,type="JKn")
svytotal(~api.stu,dJKn)

##           total    SE
## api.stu 3086009 99477

dboot = as.svrepdesign(design=dstrat,type="subbootstrap",replicates=1000)
svytotal(~api.stu,dboot)

##           total     SE
## api.stu 3086009 103337
```

c) Requerimos primero el estimador de razón y calcular el número total de estudiantes, valor último que se asume conocido

```
(r = svyratio(~api.stu,~enroll, dstrat))

## Ratio estimator: svyratio.survey.design2(~api.stu, ~enroll, dstrat)
## Ratios=
##         enroll
## api.stu  0.837
## SEs=
##          enroll
## api.stu 0.00776

tenroll = sum(apipop$enroll,na.rm=T)
```

La estimación pedida será

```
tenroll*coef(r)

## api.stu/enroll
##        3190038
```

d) Será preferible el estimador de razón, pues su error de estimación estimado es de $3\,811\,472 \times 0.007757103 = 29\,565.98$, que es casi tres veces menor que el del estimador de Horvitz-Thompson. Más aún, la estimación de razón es más cercana al verdadero número de estudiantes que tomaron el test, el cual es

```
(sum(apipop$api.stu,na.rm=T))

## [1] 3196602
```

12. Se estima que el 23.715 % de las obras están usando la metodología y el error de estimación de este porcentaje, a un nivel de confianza del 95 %, es del 5.092 %.

14. a) Construida la base muestral de datos `htree` se puede verificar lo siguiente:

```
load("/Users/lucho/Documents/TextoMuestreo2019/Texto2019_2/htree.RData")
disarb=svydesign(ids=~Lote+Num,strata=~Adm,fpc=~Nlote+Numa,nest=T,data=htree)
svymean(~Altura,disarb,deff=T)

##         mean    SE DEff
## Altura 24.77  1.17  1.8
```

b) Queda como ejercicio, pero observe que, en este caso, se tiene no un diseño estratificado por conglomerados bietápico como en a) sino uno de una sola etapa. Además, este no podrá calcularse con el paquete survey sino manualmente, pues los datos dados son solo resúmenes.
c) En la estimación de la media, el efecto de diseño se estima en 1.7988. El de b) queda como ejercicio.

16. a) Tomemos, en primer lugar, la muestra bajo el diseño propuesto

```
bb = apipop[is.na(apipop$enroll)==0,]
muestra = list()
s = c(10,5,5)
for(i in 1:3){
bbe = bb[bb$stype==levels(bb$stype)[i],]
denroll = as.numeric(by(bbe$enroll,bbe$dnum,sum))
prob = inclusionprobabilities(denroll,s[i])
set.seed(12345)
auxe = cluster(bbe,clustername=c("dnum"),s[i],method="systematic",
pik= prob,description=T)
muestra[[i]] = getdata(bbe,auxe)}
```

```
## Number of selected clusters: 10
## Number of units in the population and number of selected units: 4397 594
## Number of selected clusters: 5
## Number of units in the population and number of selected units: 751 23
## Number of selected clusters: 5
## Number of units in the population and number of selected units: 1009 12

MuestraF = do.call(rbind,muestra)
```

Las estimaciones pedidas se obtendrán con

```
disc = svydesign(ids=~dnum,strata=~stype, probs=~Prob,data=MuestraF,nest=T)
svymean(~api00,disc)

##       mean   SE
## api00  658 25.2

svyby(~api00,~stype,disc,svymean)

##   stype api00   se
## E     E   646 30.5
## H     H   581 35.5
## M     M   736 58.4

svyquantile(~api00,disc,2/3)

##       0.67
## api00  723
```

b) El análisis de regresión nos brinda el siguiente resultado:

```
rmm = svyglm(api00~emer, disc)
summary(rmm)

##
## Call:
## svyglm(formula = api00 ~ emer, disc)
##
## Survey design:
## svydesign(ids = ~dnum, strata = ~stype, probs = ~Prob, data = MuestraF,
##     nest = T)
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   732.13      30.58   23.94  5.9e-14 ***
## emer           -5.81       1.92   -3.03    0.008 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 12091)
##
## Number of Fisher Scoring iterations: 2
```

lo cuál indica que la contribución de la variable `emer` en el rendimiento de las escuelas es significativa, y se estima que por cada 1 % que se incremente el porcentaje de profesores con calificaciones de emergencia en la escuela, el rendimiento de la escuela baja en aproximadamente 5.8 puntos.

Bibliografía

Arias-Schreiber, F., Valdivieso, L. y Peña, A. (2019). *LA EVALUACIÓN DE LAS LEYES EN EL PERÚ: El análisis de costo-beneficio en el congreso de la República*, Fondo Editorial PUCP.

Bankier, M. (1988). Power allocation: Determining sample sizes for sub-national areas, *The American Statistician* **42**: 174–177.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review* **51**: 279–292.

Burnard, P. (1992). Learning from experience: Nurse tutors and student nurses perceptions of experiential learning in nurse education: Some initial findings, *International Journal of Nursing Studies* **29**: 151–161.

Cho, E. y Cho, M. (2008). The variance of sample variance from a finite population, *Survey Research Methods Section, American Statistical Association*, Denver, CO.

Cochran, W. (1977). *Sampling techniques*, Wiley Series in Probability and Statistics.

Deville, J. y Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method, *Biometrika* **85**: 89–101.

Dippo, C., Fay, R. y Morganstein, D. (1984). Computing variances from complex samples with replicate weights, *Proceedings of the Survey Research Methods Section, American Statistical Association* pp. 489–494.

Efron, B. y Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman and Hall.

Fay, R. (1984). Some properties of estimates of variance based on replication methods, *Proceedings of the Survey Research Methods Section, American Statistical Association* pp. 495–500.

Fournier, P., C. F. S. S. y Stolle, D. (2013). Canadian election study 2011: Study documentation, *Technical report*, Queen's University, Kingson, Ontario.

Fox, J. y Weisberg, S. (2018). *An R Companion to Applied Regression*, 3 edn, Sage.

Gnanadesikan, R. (1997). *Statistical data analysis of multivariate observations*, Wiley.

Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population, *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koezlemenyei* **5**: 361–374.

Hansen, M. y Hurwitz, W. (1943). On the theory of sampling from a finite population, *Annals of Mathematical Statistics* **14**: 333–362.

Heeringa, S. G., W. B. T. y Berglund, P. A. (2010). *Applied Survey Data Analysis*, Chapman and Hall.

Horvitz, D. y Thompson, D. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**: 663–685.

Khan, M.G.M., C. M. y Ahmad, N. (2006). Optimum allocation in two-stage and stratified two-stage sampling for multivariate surveys, *Proceedings of the Survey Research Methods Section, ASA* pp. 3215–3220.

Kish, L. (1965). *Survey Sampling*, Wiley Series in Probability and Statistics.

Koch, GG., F. D. y Freeman, J. (1975). Strategies in the multivariate analysis of data from complex surveys, *International Statistical Review* **43**: 59–78.

Lehtonen, R. y Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*, Jhon Wiley Sons, Ltd.

Little, R. y Rubin, D. (2002). *Statistical Analysis with Missing Data*, Jhon Wiley Sons, Inc. New Jersey.

Lohr, S. (2000). *Muestreo: Diseño y Análisis*, Internacional Thomson editores.

Lumley, T. (2010). *Complex surveys*, Wiley Series in Survey Methodology.

Lumley, T. y Scott, A. (2014). Tests for regression models fitted to survey data, *Australian and New Zealand Journal of Statistics* **56**: 1–14.

McCarthy, P. (1969). Pseudoreplication: Half-samples, *Review of the International Statistical Institute* **37**: 239–264.

Mendenhall, W., Scheaffer, R. y Ott, L. (2007). *Elementos de muestreo*, Thomson editores.

Murgia, D. (2018). Primer estudio de adpción bim en proyectos de edicicación en lima y callao 2017, *Technical report*, Pontificia Universidad Católica del Perú. Departamento de Ingeniería.

Plackett, R. y Burman, J. (1946). The design of optimum multifactorial experiments, *Biometrika* **33**: 305–325.

Quenouille, M. H. (1949). Approximate tests of correlation in time series, *Journal of the Royal Statistical Society B* **11**: 68–84.

Rao, J. y Scott, A. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables, *Journal of the American Statistical Association* **76**: 221–230.

Rao, J. y Scott, A. (1984). On chi-squared tests for multiway contigency tables with proportions estimated from survey data, *Annals of Statistic* **12**: 46–60.

Rao, J. y Wu, C. (1988). Resampling inference with complex survey data, *Journal of the American Statistical Association* **83**: 231–241.

Richardson, M. (2012). Sampling in archeology, *STatistics Education Web*. pp. 1–18.

Sampford, M. (1967). On sampling without replacement with unequal probabilities of selection, *Biometrika* **54**: 499–513.

Satterthwaite, F. (1946). *An approximate distribution of estimates of variance components*, Biometrics Bulletin, 2 110-114.

Thomas, D. y Rao, J. (1990). Small-sample comparison of level and power for simple goodness-of-fit statistics under cluster sampling, *Journal of the American Statistical Association* **82**: 630–636.

Thomas, D.R., S. A. y Roberts, G. (1996). Tests of independence on two- way tables under cluster sampling: An evaluation, *International Statistical Review* **64**: 295–311.

Tillé, I. (2006). *Sampling Algorithms*, Springer.

Tukey, J. W. (1958). Bias and confidence in not quite large samples, *Annals of Mathematical Statistics* **29**: 614.

Valdivieso, L. (2017). *Estadística aplicada. Notas de clase*, PUCP.

Valliant, R. (1993). Post-stratification and conditional variance estimation, *JASA* **88**: 89–96.

Valliant, R., Dever, J. y Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*, Springer.

Wolter, K. (2007). *Introduction to Variance Estimation*, Springer.

Ypma, J., Borchers, H. y Eddelbuettel, D. (2018). *nloptr: R Interface to NLopt*, R package version 1.2.1.
URL: *https://CRAN.R-project.org/package=nloptr*

La gran mayoría de las investigaciones trabajan con datos, los cuales pueden obtenerse a través de la observación de una o más variables en una población o muestra. Si bien una muestra puede ser cualquier conjunto de una población, conclusiones válidas sobre esta última sólo podrán garantizarse de ser la muestra probabilística, es decir, en las que cada unidad seleccionada tenga una probabilidad conocida de ser tomada. Este texto introduce las principales técnicas para seleccionar y analizar este tipo de muestras cuando la población es finita. La finitud es aquí relevante, pues hace que el desarrollo de estas técnicas se oriente más por un enfoque basado en el diseño. En él, la aleatoriedad de los resultados es producto del proceso de selección de la muestra y no de la consideración de que la o las variables de interés provienen de un hipotético modelo poblacional como se acostumbra asumir en la inferencia clásica. Aparte de las técnicas o esquemas de muestreo básicos como el del muestreo aleatorio simple, el muestreo estratificado y el de conglomerados, el texto introduce algunos tópicos de muestreo complejo. Este, que en la práctica es el esquema más utilizado, se origina cuando debido a las restricciones presupuestales y logísticas o la configuración y tamaño de la población se hace necesario el restringir o combinar dos o más esquemas básicos ya sea que las selecciones se hagan con igual probabilidad o no. Parte central y transversal del desarrollo del texto será el uso del software libre R, con principalmente los paquetes survey y sampling. El texto incluye también varios ejercicios propuestos y soluciones o sugerencias a todos los problemas pares. Muchos de los ejemplos desarrollados en el texto y de los ejercicios planteados se basan en datos reales locales o foráneos de dominio público.

ISBN: 978-612-47757-1-0